Pelvic Tilt in Sitting: Do You See What I See? (Maybe Not)

Matthew K. Bagg *PhD*
Post-Doctoral Research Fellow, Centre for Pain IMPACT, Neuroscience Research Australia; Curtin Health Innovation Research Institute, Faculty of Health Sciences, Curtin University; Perron Institute for Neurological and Translational Science, Perth, Australia

Dr Ian Skinner *PhD*
Senior Lecturer, Associate Head of School, Physiotherapy; School of Community Health, Faculty of Science, Charles Sturt University, Port Macquarie, New South Wales, Australia

Niamh Moloney *PhD*
Associate Professor, Department of Exercise Sciences, University of Auckland, Auckland, New Zealand

Martin Lock *BHSc (Physiotherapy)*
Lead Physiotherapist, Persistent Pain, Guernsey Therapy Group, Guernsey

James McAuley *PhD*
Senior Research Scientist, Director, Centre for Pain IMPACT, Neuroscience Research Australia; Professor, Faculty of Medicine and Health, University of New South Wales; Honorary Research Fellow, The George Institute for Global Health, Australia

Martin Rabey *PhD*
Adjunct Research Fellow, School of Allied Health, Curtin University, Perth, Australia

ABSTRACT

Examination of pelvic tilt movements are utilised across many fields of physiotherapy. It is important for physiotherapists to establish a clinically helpful, time-efficient test assessing pelvic tilt, reliable within and across multiple assessors. Elgueta-Cancino et al. (2014) described such a test; however, their methodology reduced clinical applicability and revealed limitations regarding examination of test reliability. This study aimed to independently evaluate the reliability of a clinical test of pelvic tilt. Twenty-three participants with chronic low back pain completed the test following standardised instructions and demonstration by one assessor. Participants tilted the pelvis forwards and backwards 10 times in sitting. The test was simultaneously scored on the scale originally described by three blinded assessors. Participants repeated the test one-week later. Inter-assessor reliability was determined using an intra-class correlation coefficient (ICC 2,1), with a resulting value of 0.52, 95% confidence interval [0.35–0.68]; and a standard error of measurement SEM (with a resulting value of 1.28). The following SEM values were found for intra-assessor agreement: Assessor 1 =1.52, assessor 2 =1.47, and assessor 3 = 1.19. These findings suggest the inter- and intra-assessor reliability of a clinical test of pelvic tilting has insufficient reliability to distinguish between participants across multiple assessors. An observed change of at least 1.5 points may be necessary to be confident true change in test performance has occurred.

Bagg, M. K., Skinner, I., Moloney, M., Lock, M., McAuley, J., & Rabey, M. (2023). Pelvic tilt in sitting: Do you see what I see? (Maybe not). *New Zealand Journal of Physiotherapy, 51*(1), 48–52. https://doi.org/10.15619/NZJP/51.1.06

Key Words: Low Back Pain, Movement Control, Reliability

INTRODUCTION

Many methods of examining lumbopelvic movement patterns, particularly in relation to low back pain, are described in the physiotherapy literature. However, examination of a person's ability to perform pelvic tilting, and subsequent rehabilitation of this movement, is utilised across many fields of physiotherapy – for example, musculoskeletal (Elgueta-Cancino et al., 2014), respiratory (Aramaki et al., 2021), continence (Berghmans et al., 2020), and neurology (Karthikbabu et al., 2017). In the research setting, pelvic tilt is commonly examined using electromyography and kinaesthetics (Dankaerts & O'Sullivan, 2011), which is expensive and impractical clinically. Therefore, it is important for physiotherapy practice to establish a clinically helpful test to assess pelvic tilt, which should be time-efficient and reliable both within and across multiple assessors.

Movement patterns, for example in people with chronic low back pain (CLBP) (Dankaerts & O'Sullivan, 2011; Hodges & Smeets, 2015) are complex. Therefore, even for a movement as seemingly simple as pelvic tilting, physiotherapists must consider factors including range of movement, localisation of the movement, muscular control of the movement, and concurrent respiratory pattern. A valid and reliable test incorporating such factors is important to facilitate practice across many fields of physiotherapy and communication between therapists. Elgueta-Cancino et al. (2014) describe a potentially comprehensive, time-efficient clinical test of pelvic tilting in sitting. The participants watched a standardised instruction video including a demonstration and verbal instructions to tilt the pelvis anteriorly and posteriorly 10 times, followed by 2 min supervised training of the movement. Subsequently, to standardise the movement examination, the assessor used a scale covering quality (smoothness, range) of

pelvic movement, control of adjacent regions (thoracolumbar movement, erector spinae activity), directional influence on movement quality, ability to breathe during movement, and ability to perform quality movements repeatedly. A total score was derived, ranging 0–10 points, with higher scores reflecting greater movement control. However, while use of the scale appears time-efficient, the training process participants completed may be impractical in a clinical setting.

Adequate inter- and intra-assessor reliability is important for the validity of clinical tests (Dankaerts et al., 2006). Elgueta-Cancino et al. (2014) report the inter- and intra-assessor reliability of their test of pelvic tilting to be substantial/moderate. However, intra-assessor reliability was examined with a single assessor and inter-assessor reliability with only two assessors. Whilst the reported kappa values might be interpreted as moderate (0.15–0.66), confidence intervals were large and deteriorated after training. The reliability of this test has also yet to be replicated independently.

Therefore, the aim of this study was to independently evaluate the reliability of a clinically applicable test of pelvic tilting across multiple assessors at two time-points in people with CLBP.

METHODS

A test-retest design was implemented, with participants rated by three assessors at two time-points, one-week apart. People with CLBP were recruited from the public via multimedia advertisements. We used an interval estimation to prospectively calculate sample size using the R package "presize" (Lenz & Haynes, 2020; R Core Team, 2020). Twenty-three participants were required to detect an intra-class correlation coefficient (ICC) of 0.85 with three assessors and a desired confidence interval of 0.2 with 95% confidence (Bonett, 2002). This research received approval from the Guernsey Ethics Committee (approval number IJG/C5.4) and complied with the Declaration of Helsinki (World Medical Association, 2013). Participants gave informed written consent.

Participants

Potential participants contacted researchers and were screened to determine compliance with inclusion (18–70 years old; CLBP > 3-months duration, with or without leg pain) and exclusion criteria (serious spinal pathology such as cancer or inflammatory arthropathy, diagnosed neurological conditions, clinically determined nerve root compromise, and pregnancy).

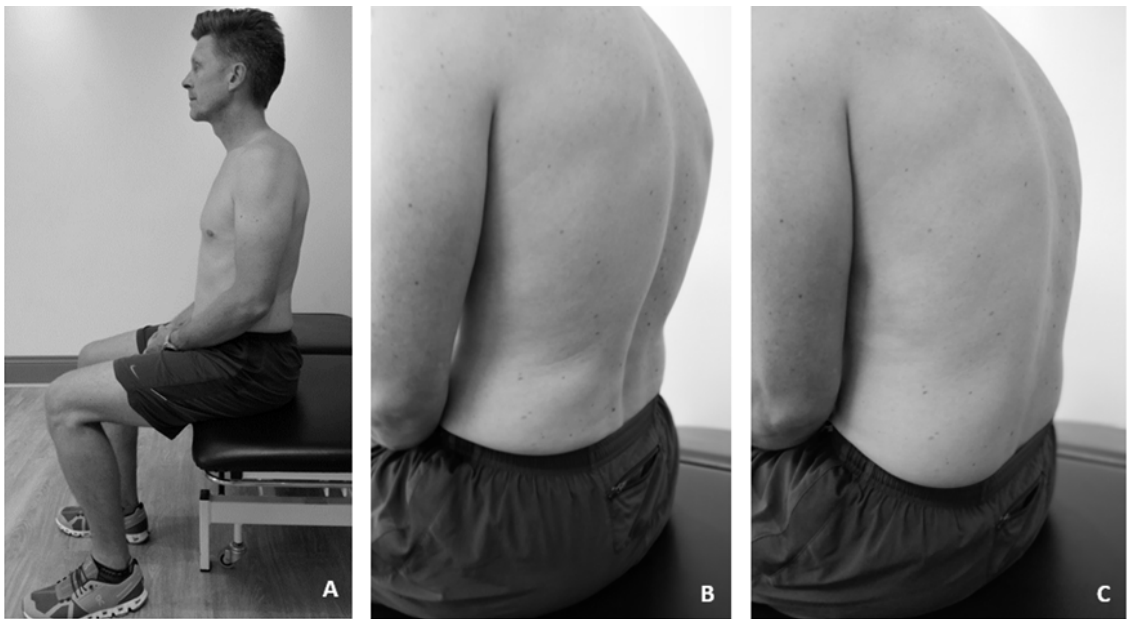
Testing procedure

Three physiotherapists were assessors (MR, NM, ML). Two assessors had 20 and 22 years of clinical experience, respectively, and Master's and PhD degrees in musculoskeletal pain/physiotherapy. The third had 13 years clinical experience. Assessors completed one 30 min preparatory session together on demonstrating the test to participants and familiarisation and standardisation of scoring.

Participants completed the following protocol for the clinical test of lumbopelvic control: Standardised verbal instructions, and demonstration of performance of the test were given by one assessor (randomly selected) using wording described by Elgueta-Cancino et al. (2014). Participants were seated on an adjustable height plinth so that both hips and knees were at approximately 90° of flexion, with the feet flat on the floor. The test involves tilting the pelvis forwards and backwards 10 times in sitting (Figure 1). All assessors concurrently watched the participant perform the test and scored the participant's performance on the scale described by Elgueta-Cancino et al. The scale includes scores for different movement components: quality (smoothness, range) of pelvic movement (0–3 points), control of adjacent regions (thoracolumbar movement, erector spinae activity) (0–3 points), directional influence on movement

Figure 1.

Clinical Test of Lumbopelvic Control



Note. Images showing the test position in sitting (panel A). The test involves anterior (panel B) and posterior (panel C) pelvic tilting, 10 repetitions.

quality (0–2 points), ability to breathe during movement (0–1 point), and ability to perform quality movements repeatedly between (0–1 point). The total score ranges between 0–10 points with higher scores reflecting greater movement control. Assessors were blinded to each other's scores.

Participants were instructed not to practise the movement and returned one week later to repeat the test. The verbal instructions, demonstration, and scoring procedures were repeated.

Data analyses

Data supporting the findings of this study were uploaded to the Open Science Framework (https://osf.io/) and are available from the corresponding author. Data are not publicly available due to ethical restrictions.

Inter-assessor reliability, inter-assessor agreement, and intra-assessor agreement were calculated using *total* scores for each participant. We did not evaluate reliability or agreement of *individual* items because we were interested in the overall test format in clinical use.

Inter-assessor reliability was calculated with an ICC (2,1) (Shrout & Fleiss, 1979) using a two-way random effect model with absolute agreement, using a single measurement (McGraw & Wong, 1996). The ICC provides a measure of relative reliability indicating the similarity of scores between two measurements, relative to the overall distribution of scores (Scholtes et al., 2011). ICC scores are comparable to the kappa values used by Elgueta-Cancino et al. (2014) but with the advantage of considering systematic differences between assessors and extending generalisability of scores to other assessors (Streiner et al., 2014). We considered an ICC of 0.7 indicative of sufficient inter-assessor reliability (Nunnally & Bernstein, 1994), in keeping with recommendations not to use arbitrary classification systems for interpretation of reliability coefficients (de Vet et al., 2011; Streiner et al., 2014).

Standard error of measurement (SEM) was calculated to assess inter- and intra-assessor agreement. The SEM provides a value, in the unit of measurement of the test, of the absolute difference in scores. We calculated the SEM as the square root of the error variance $\sqrt{\sigma^2_{error}}$ (de Vet et al., 2006). We accounted for systematic differences between assessors and testing sessions by including in the error variance both the residual variance ($\sigma^2_{residual}$) (and either the (i) assessor variance (σ^2_{pt}) or (ii) the session variance ($\sigma^2_{session}$), depending on whether (i) inter-assessor or (ii) intra-assessor SEM was being calculated (de Vet et al., 2006). Variance components were estimated in STATA (StataCorp. 2017. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC.), using a random effects model fit with restricted maximum likelihood and participants' score as the dependent variable. There are no strict criteria for evaluating minimum thresholds for SEM values. Values should be interpreted with reference to the context in which the measurement instrument is applied.

The SEM value for inter-assessor agreement provides information on the consistency between scores from different assessors of the same participant (Weir, 2005). A low SEM value is preferable. We calculated the SEM for inter-assessor reliability for the three assessors from both testing sessions, using the formula $\sqrt{(\sigma^2_{pt} + \sigma^2_{residual})}$ (de Vet et al., 2006). Participants and assessors were considered factor variables when estimating variance components. Data from both testing sessions were used and each testing session was considered an independent sample. We calculated the mean score and standard deviation for each assessor across all observations to provide perspectives of both time points.

The SEM value for intra-assessor agreement provides information on consistency between scores from the same assessor at repeat assessments of the same participant (Weir, 2005). A low SEM value is preferable. The intra-assessor agreement indicates the sensitivity of the tool to be used in an *evaluative* (longitudinal) manner, such as observing the effect of an intervention on lumbopelvic control. We calculated the SEM for intra-assessor agreement for all three assessors across both sessions, using the formula $\sqrt{(\sigma^2_{session} + \sigma^2_{residual})}$ (de Vet et al., 2006). Participants and testing sessions were considered factor variables when estimating variance components.

RESULTS

We recruited 23 participants (69.6% female, mean age 55.4 years; range 23–68 years) who attended both testing sessions.

The inter-assessor reliability of the clinical test of lumbopelvic control was ICC (2,1) = 0.52, 95% CI [35, 0.68]. The inter-assessor agreement of the test was SEM = 1.28. Table 1 contains mean scores, standard deviation, and variance values for the three assessors.

Intra-assessor agreement values were: assessor 1 SEM = 1.52, assessor 1 SEM = 1.47, assessor 3 SEM = 1.19. Table 2 contains mean scores, standard deviation, and variance values for sessions 1 and 2 for each assessor.

Table 1

Mean Scores, Standard Deviations, and Variance Values Used to Calculate Inter-Assessor Reliability and Inter-Assessor Agreement (n = 46)

Assessor	Mean score (0–10 points)	*SD* (0–10 points)	Participant variance	Assessor variance	Residual variance
1	3.52	1.92	1.78	6×10^{-2}	1.57
2	3.79	1.98			
3	3.17	1.57			

Table 2

Mean Scores, Standard Deviations, and Variance Values Used to Calculate Intra-Assessor Agreement (n = *23)*

Assessor	Session 1		Session 2		Participant variance	Session variance	Residual variance
	M	SD	*M*	SD			
1	3.59	2.05	3.46	1.83	1.40	4.31×10^{-18}	2.32
2	4.09	2.19	3.50	1.74	1.84	8.3×10^{-2}	2.06
3	3.26	1.69	3.09	1.47	1.06	9.19×10^{-17}	1.42

DISCUSSION

We independently evaluated the reliability and agreement of a clinical test of lumbopelvic control across multiple assessors at two time-points. Our results suggest that when the test is administered by multiple assessors there is considerable variance in scores not due to a true difference among participants. Therefore, the test may not distinguish between participants due to the comparatively higher variance of assessors and random variance in the test itself (ICC for inter-assessor reliability (2,1) = 0.52, 95% CI [0.35, 0.68] (Table 1). The upper bound (0.68) of the 95% CI does not meet the minimum criterion of 0.7 and the lower bound (0.35) is well short. The SEM for inter-assessor agreement indicates that if an assessment of the same person is made by multiple assessors, scores may vary by 1.28 points on the 0–10 scale. The SEM values for intra-assessor agreement ranged from 1.19 to 1.52, suggesting repeated assessments by the same assessor require that observations differ by at least 1.52 points to demonstrate change not attributable to measurement error.

SEM values for inter-assessor agreement can be used to interpret ICC values for inter-assessor reliability. ICC values indicate similarity of scores between participants relative to the overall spread of scores. The overall spread should be sufficient to adequately distinguish participants. The ICC will be low when this does not occur, even if assessors give similar scores (there is good consistency). Sufficient spread is judged using the standard deviation of scores and SEM. The standard deviation ranged from 1.57 to 1.98 (Table 1) – a small spread – indicating most participants scored within 2 points of one another. The SEM indicates scores varied by 1.28 points between assessors. Together, these values indicate insufficient spread to distinguish participants. The spread of scores is not much greater than the observed variability between assessors. This may have contributed to the low ICC values observed. Future evaluations of this test might consider adapting the scale to allow greater spread of scores.

Our results differ with those previously reported. Elgueta-Cancino et al. (2014) evaluated inter-assessor reliability using Cohen's kappa across two assessors and did not calculate agreement. We evaluated inter-assessor reliability with an ICC across three assessors and calculated agreement. Our result may be more robust because we evaluated three assessors and used a larger sample. Our results may have greater interpretability and clinical application because ICCs are more generalisable measures of inter-assessor reliability than Cohen's kappa (de Vet et al., 2011). Second, values for agreement are expressed on the test scale.

Elgueta-Cancino et al. (2014) evaluated intra-assessor reliability for a single assessor of 10 participants on two occasions. Participants were assessed *in vivo* on the first occasion and the assessor reviewed a video taken of that same performance on the second occasion. We evaluated intra-assessor reliability for three assessors of 23 participants at two time-points, under identical conditions *in vivo*. This more closely reflects clinical testing.

Our results may also differ because participants received less training than the study by Elgueta-Cancino et al. (2014). We did not train participants beyond standardised instructions and demonstration of the test (duration < 60 s). Whereas, Elgueta-Cancino et al. (2014) provided initial training using a video and 2 min of training following the first test performance. There may be an effect of training on test performance, although this is uncorroborated. Interestingly, inter-and intra-assessor reliability reduced from substantial to moderate after 2 min of training (Elgueta-Cancino et al., 2014). Regardless, the demonstration used in this study likely more closely reflects use of the test clinically.

Our work is robust in several respects. We prospectively calculated sample size for a broader number of measures of reliability. We employed three assessors, with broad experience, and conducted tests in clinically representative conditions *in vivo*. We prospectively registered the Statistical Analysis Plan and our data and analytic code are available upon request.

Unfortunately, limited data on participant characteristics complicate comparison with other studies. As potential change in participant's presentations was not considered, it is possible their ability to perform the test differed across time-points, adversely influencing examination of test reliability. In addition, we assumed that the total scores used to assess the SEM and ICC are continuous, an assumption generally accepted as necessary for using the SEM. An argument could be made that the total scores are not continuous, which should be considered. However recent evidence has indicated that ICC and SEMs may still be appropriate if the data is not continuous (de Raadt et al., 2021).

CONCLUSION

The clinical testing of lumbopelvic control is time-efficient and involves functional movement that can be used within rehabilitation. However, our results question the reliability of the

test. Examination of other tests may reveal an alternative test that is reliable. Conversely, it may be that more complex clinical movement examination processes or technological movement assessment equipment are necessary to capture lumbopelvic movement control reliably.

KEY POINTS

1. Inter- and intra-assessor reliability of a clinical test of pelvic tilting has insufficient reliability to distinguish between participants across multiple assessors.
2. An observed change of at least 1.5 points may be necessary to be confident true change in test performance has occurred.
3. Physiotherapists may need to consider other tests, complex clinical movement examination processes, or technological movement assessment equipment to capture lumbopelvic movement control reliably.

DISCLOSURES

No funding was obtained for the study. MB was supported by a NeuRA PhD Candidature Scholarship and Supplementary Scholarship and was supported during this work by an Australian Research Training Program Scholarship and a University of New South Wales Research Excellence Award. MB has received conference travel support from the Chiropractor's Association of Australia and Memorial University of Newfoundland to speak about unrelated topics. The other authors have no conflicts of interest to declare.

PERMISSIONS

This research received approval from the Guernsey Ethics Committee (approval number IJG/C5.4) and complied with the Declaration of Helsinki (World Medical Association, 2013). Participants gave informed written consent. The photographs in Figure 1 are of one of the authors, who provided permission for publication.

CONTRIBUTIONS OF AUTHORS

MB was involved in conception of the research idea, literature review, data analysis, interpretation and writing and review of the final manuscript. IS was involved in data analysis, interpretation and writing and review of the final manuscript. NM and ML were involved in data collection and writing and review of the final manuscript. JM was involved in conception of the research idea and writing and review of the final manuscript. MR was involved in conception of the research idea, literature review and writing and review of the final manuscript.

ADDRESS FOR CORRESPONDENCE

Martin Rabey, 1/13 Garden Terrace, Devonport, Auckland 0624, New Zealand.

Email: martinrabey@gmail.com

REFERENCES

Aramaki, Y., Kakizaki, F., Kawata, S., Omotehara, T., & Itoh, M. (2021). Effects of the posterior pelvic tilt sitting posture on thoracic morphology and respiratory function. *Journal of Physical Therapy Science*, *33*(2), 118–124. https://doi.org/10.1589/jpts.33.118

Berghmans, B., Seleme, M. R., & Bernards, A. T. M. (2020). Physiotherapy assessment for female urinary incontinence. *International Urogynecology Journal*, *31*(5), 917–931. https://doi.org/10.1007/s00192-020-04251-2

Bonett, D. G. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine*, *21*, 1331–1335. https://doi.org/10.1002/sim.1108

Dankaerts, W., & O'Sullivan, P. (2011). The validity of O'Sullivan's classification system (CS) for a sub–group of NS-CLBP with motor control impairment (MCI): Overview of a series of studies and review of the literature. *Manual Therapy*, *16*(1), 9–14. https://doi.org/10.1016/j.math.2010.10.006

Dankaerts, W., O'Sullivan, P. B., Straker, L. M., Burnett, A. F., & Skouen, J. S. (2006). The inter-examiner reliability of a classification method for non-specific chronic low back pain patients with motor control impairment. *Manual Therapy*, *11*(1), 28–39. https://doi.org/10.1016/j.math.2005.02.001

de Raadt, A., Warrens, M., Bosker, R., & Kiers, H. (2021). A comparison of reliability coefficients for ordinal rating scales. *Journal of Classification, 38*(3), 519–543. https://doi.org/10.1007/s00357-021-09386-5

de Vet, H., Terwee, C., Knol, D., & Bouter, L. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology*, *59*, 1033–1039. https://doi.org/10.1016/j.jclinepi.2005.10.015

de Vet, H., Terwee, C., Mokkink, L., & Knol, D. (2011). *Measurement in medicine: A practical guide.* Cambridge University Press.

Elgueta-Cancino, E., Schabrun, S., Danneels, L., & Hodges, P. (2014). A clinical test of lumbopelvic control: Development and reliability of a clinical test of dissociation of lumbopelvic and thoracolumbar motion. *Manual Therapy*, *19*(5), 418–424. https://doi.org/10.1016/j.math.2014.03.009

Haynes, A. G., Lenz, A., Stalder, O. & Limacher, A. (2021). presize: An R-package for precision-based sample size calculation in clinical research. *Journal of Open Source Software, 6*(6), 3318. https://doi.org/10.21105/joss.03118

Hodges, P. W., & Smeets, R. J. (2015). Interaction between pain, movement, and physical activity: Short-term benefits, long-term consequences, and targets for treatment. *Clinical Journal of Pain*, *31*(2), 97–107. https://doi.org/10.1097/AJP.0000000000000098

Karthikbabu, S., Chakrapani, M., Ganesan, S., & Ellajosyla, R. (2017). Pelvic alignment in standing, and its relationship with trunk control and motor recovery of lower limb after stroke. *Neurology and Clinical Neuroscience*, *5*(1), 22–28. https://doi.org/10.1111/ncn3.12092

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

R Core Team. (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Scholtes V. A., Terwee C. B., & Poolman R. W. (2011). What makes a measurement instrument valid and reliable? *Injury, 42(3)*, 236–240. https://doi.org/10/d7264j

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86(2)*, 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Streiner, D. L., Norman, G. R., & Cairney, J. (2014). *Health measurement scales: A practical guide to their development and use* (5th ed.). Oxford University Press.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, *19(1)*, 231–240. https://doi.org/10.1519/15184.1

World Medical Association. (2013). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *Journal of the American Medical Association*, *310*(20), 2191–2194. https://doi.org/10.1001/jama.2013.281053