Moderate reliability of the lateral step down test amongst experienced and novice physical therapists

Cody Mansfield, PT, DPT, ATC, OCS\textsuperscript{a,b,c}, Claire Spech, BS\textsuperscript{a}, Katherine Rethman, PT, DPT, SCS\textsuperscript{b,c}, Sarah Clagg, PT, DPT, ATC, SCS\textsuperscript{b}, Adam Ingle, PT, DPT, SCS\textsuperscript{b}, Adrian Largent, BS\textsuperscript{b}, Thanvi Vatti, BA\textsuperscript{d}, Matt Morrow, BS\textsuperscript{e}, Lucas VanEtten, PT, DPT, OCS\textsuperscript{b}, Matthew Briggs, PT, DPT, ATC, PhD, SCS\textsuperscript{b,c,f}

\textsuperscript{a}School of Health and Rehabilitation Sciences, College of Medicine, Ohio State University, Columbus, OH, USA
\textsuperscript{b}OSU Sports Medicine, Ohio State University Wexner Medical Center, Columbus, OH, USA
\textsuperscript{c}Sports Medicine Research Institute, Ohio State University Wexner Medical Center, Columbus, OH, USA
\textsuperscript{d}College of Arts and Sciences, Case Western Reserve University, Cleveland, OH, USA
\textsuperscript{e}Department of Neuroscience, College of Medicine, Ohio State University, Columbus, OH, USA
\textsuperscript{f}Department of Orthopaedics, Ohio State University, Wexner Medical Center, Columbus, OH, USA

Abstract

\textbf{Background/Introduction:} The lateral step-down test is used to appraise movement quality in patients with patellofemoral pain (PFP), however, it is unclear if reliability of the test is affected by physical therapist experience.

\textbf{Objective:} Determine if there is a difference in reliability between ‘experienced’ and ‘novice’ physical therapists appraising movement quality of patients with PFP during the lateral step-down test.

\textbf{Methods:} Three ‘experienced’ and 3 ‘novice’ physical therapists analyzed movement quality of 22 participants [mean age (SD) 28.25 (6.5) years] with PFP. Physical therapists viewed two-dimensional videos of participants performing the lateral step-down test and appraised the quality with a score (0–1 = ‘good’; 2–3 = ‘fair’, and 4–5 = ‘poor’) at baseline and 1 week. Inter- and intra-rater reliability were calculated with kappa and percent agreement. Differences between the groups were assessed with the chi-square test with an a priori alpha level of < 0.05.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

\textbf{CONTACT} Cody Mansfield, Cody.mansfield@osumc.edu, School of Health and Rehabilitation Sciences, College of Medicine, Ohio State University, 453 W. 10th Avenue, Columbus, OH 43210.

\textbf{Declaration of Interest} The authors declare no conflict of interest.
**Results:** Inter- and intra-rater reliability ranged from fair to moderate ($\kappa = 0.40–0.65$). There was no difference in reliability between ‘experienced’ and ‘novice’ physical therapists at baseline ($p = .13$) or 1 week post testing ($p = .94$).

**Conclusions:** There was no difference in reliability between ‘experienced’ and ‘novice’ physical therapists using categories to appraise movement quality during the lateral step-down test for patients with chronic PFP.

**Keywords**
Kinesiology; knee; motion; physiotherapy; reliability; video analysis

**Introduction**

The lateral step down test (LSDT) is used to appraise movement quality at the trunk, pelvis, and knee in patients with patellofemoral pain (PFP) (Piva et al., 2009; Silva et al., 2019; Willy et al., 2019). To complete the test, a patient will stand on top of a step with the foot placed along the medial edge and perform a single leg step down multiple times without allowing their weight to shift onto their contralateral extremity, while a physical therapist makes note of the number of movement deviations deemed suboptimal (Figure 1) (Silva et al., 2019). The cumulative number of deviations during the LSDT determines the quality category assigned: ‘good’ movement with 0–1 deviations, ‘fair’ movement with 2–3 deviations, and ‘poor’ movement with 4 or more deviations (Piva et al., 2006) (Table 1). Despite the wide use of the LSDT, it has only been studied in patients with acute and subacute PFP (<3 months), and differences in reliability between novice and experienced physical therapists remains unknown (Silva et al., 2019).

Reliability of the LSDT in patients with chronic PFP (≥3 months) has not been studied (Silva et al., 2019). Only 4 studies have investigated the reliability of categorizing movement (Chmielewski et al., 2007; Piva et al., 2006; Rabin and Kozol, 2010; Rabin et al., 2014). In a population with PFP for at least 4 weeks, substantial reliability ($\kappa = 0.67$) was achieved using a 3 category system (‘good’, ‘fair’, or ‘poor’) to appraise movement quality (Piva et al., 2006), whereas high/almost perfect reliability ($\kappa = 0.81$) occurred with a 2 category system (‘good’, ‘moderate’) in a homogenous population of young Israel Defense Soldiers (Rabin et al., 2014). Two other studies investigated the reliability in a healthy population and yielded low ($\kappa = 0.19$) (Chmielewski et al., 2007) to moderate ($\kappa = 0.59$) (Rabin and Kozol, 2010) reliability scores.

It is unknown if reliability of the LSDT is impacted by the experience level of a physical therapist. The majority of studies investigating the LSDT used 2 to 4 physical therapist or athletic trainer raters, and yielded moderate to high reliability (Piva et al., 2006; Rabin and Kozol, 2010; Rabin et al., 2014). A study with 3 raters yielded the lowest reliability score ($\kappa = 0.19$) (Chmielewski et al., 2007). In these studies, years of experience and clinical expertise were not considered in the reliability of the LSDT. Differences in patient reported outcomes between physical therapists with specialty board certifications, years of experience, residency and fellowship training have been identified (Rodeghero et al., 2015). More studies need to consider years’ experience and/or previous training in the reliability.
of tests commonly used in research and clinical practice. The objective of our study was to determine if there is a difference in reliability between ‘experienced’ and ‘novice’ physical therapists appraising movement quality of patients with PFP during the LSDT. We hypothesized that the LSDT would yield moderate to high reliability scores amongst physical therapists evaluating patients with chronic PFP, and that physical therapists with more years of experience, who are board certified specialists will have higher inter- and intra-rater reliability than ‘novice’ physical therapists without board specialty certification.

Methods

Participants

**Raters**—We recruited 3 ‘experienced’ physical therapists and 3 ‘novice’ physical therapists based on years’ experience and post-professional specialty certifications from the American Board of Physical Therapy Specialties (American Board of Physical Therapy Specialties, 2017). Five years’ experience and at least 1 post-professional specialty board certification (orthopedic certified specialist or sports certified specialist) suggestive of obtainment of advanced clinical knowledge in the domains of orthopedics or sports physical therapy (Rodeghero et al., 2015) was the cut off to be in the ‘experienced’ group. All other physical therapists were placed in the ‘novice’ group. Table 2 describes pertinent rater characteristics. All physical therapists included in the study practiced in high volume outpatient orthopedic/sports clinical settings with patients with knee pain constituting the highest percentage of patients evaluated and treated.

**Participants with patellofemoral pain and performance of lateral step down test**—Two-dimensional (2D) videos of 22 participants [14 females and 8 males; mean age (SD) 28.25 (6.5) years] with PFP were retrospectively viewed by the physical therapist raters (Table 3). All participants had unilateral PFP for greater than 3 months, with pain around the patella during squatting, running, or ascending/deadening stairs and no signs of other conditions that could cause anterior knee pain (Willy et al., 2019). The participants were referred to an outpatient physical therapy facility by a physician with diagnosis related to PFP and no evidence of osteoarthritis or other form of knee derangement or history of patellar instability (Willy et al., 2019). All participants reported PFP around the patella during ascending/descending stairs, squatting, and/or running (Nakagawa, Maciel, and Serrão, 2015; Nakagawa, Serrão, Maciel, and Powers, 2013). Participants were excluded from the study if they had signs of knee osteoarthritis, bilateral PFP, had surgery to the knee within 6 months, or age outside the range of 18 to 40 years old. Participants were asked to perform the LSDT on their painful limb for 5 repetitions. The step was standardized at 20 cm and not adjusted for height of the participant. Participants were allowed to practice and warm up prior to executing the 5 repetitions.

Two GoPro cameras (GoPro HERO 3; GoPro, San Mateo, CA) were used on tripods to simultaneously capture the LSDT in the frontal and sagittal planes during all 5 repetitions. The resolution of the cameras was set to 1080p at 120 frames per second. The front camera was placed 96 inches from the toe of the stance foot, and the side camera was placed 111 inches directly lateral from the outside of the knee. The camera height was set
approximately in line with the participants’ knee center, but with full view of the body of the participant.

**Design**—The project and study protocol were reviewed and permission to conduct the study was granted by the Institutional Review Board. Approved informed consents were obtained from all participants. The raters watched the 2D video clip of each participant performing the LSDT and were asked to rate the overall quality of movement based on any deviations from normal movement during any of the repetitions. The raters selected deviations from Table 1 if observed during the LSDT (Rabin et al., 2014). The scoring criteria assigns a score of 1 to each deviation except for tibial tuberosity movement medial to medial border of foot, which is assigned a score of 2 (Chmielewski et al., 2007; Jones et al., 2014; McGovern, Martin, Christoforetti, and Kivlan, 2018; Piva et al., 2009, 2006; Rabin and Kozol, 2010; Rabin et al., 2014; Rabin, Portnoy, and Kozol, 2016). The researcher instructed the raters to identify the number of deviations observed in all 5 repetitions of the LSDT. Raters in this study were asked to select the specific deviations observed from each participant performing the LSDT cumulatively for all five repetitions and were processed post-testing to determine the quality of movement for data analysis.

**Physical therapy rating procedures and assessment of the lateral step sown test**

The captured frontal and sagittal plane 2D video from each of the participants were merged together for the rater to simultaneously view 5 repetitions of the LSDT on a computer screen from both planes. This 2D video from each participant was then compiled into 1 continuous video clip to be viewed by the physical therapist raters. Raters were not allowed to pause or slow down the video while watching a participant complete the five repetitions but could pause the video in between participants. A unique identifier was assigned to each participant which then appeared on the screen in the video clip before their LSDT began. Raters viewed the videos of participants in order of when the participant entered the study.

Data collection took place in the computer lab of the research facility with a dual monitor setup. One monitor was used for viewing of the video and the other for data entry. Data was entered and managed using REDCap™ (Research Electronic Data Capture), a secure, web-based application used for data capture, auditing, and exporting with features to gather rater details, describe study protocol, and display appropriate choices for rating movement (Harris et al., 2019, 2009). Each physical therapist rater viewed each participant’s LSDT 1 time and then viewed and rated the video again 1 week later to establish inter- and intra-rater reliability. Raters were blinded to their previous scores and scores of others. The same identifying numbers were used for each participant with PFP and the videos were presented in the same order from each testing session.

Data collection for each rater was independently conducted and lasted approximately 30 min. Prior to starting data collection for this study, each rater reviewed an instructional guide, designed by the investigators of the study. The guide included information on the LSDT, scoring criteria (Table 1), and instructions on how to enter data into the REDCap™ program (Harris et al., 2019). The rater reviewed the testing protocol with the investigator and clarified any questions on LSDT rating before launching REDCap™. REDCap™ was
set up a survey with a specialized link sent to each rater. First, the rater entered demographic information followed by details of their professional experience and credentials. Secondly, the rater reviewed the looped 2D video of all 22 participants performing the LSDT on 1 monitor and simultaneously entered their assessment into the REDCap™ program on another monitor by selecting any movement deviations from a displayed list of the 6 deviations (Table 1).

It was emphasized to the raters that they were unable to pause or slow down the video of a participant performing the LSDT but were allowed to pause the video in between participants in order to enter the data onto REDCap™. They were not allowed to rewind or watch the video again for any of the participants. The number of repetitions was the same as others (Piva et al., 2006), where reliability was determined during live viewing of participants performing the test. We chose not to allow raters to rewind to better mimic a live viewing of a testing session. One study member (TV) was in charge of supervising all raters during the testing. All testing for each rater was conducted at the same site with the same investigator (TV) providing instruction and ensuring all data was properly entered.

**Statistical analysis**

Inter- and intra-rater reliability were calculated with kappa and percent agreement. The kappa values were interpreted based on a scale that deemed values of 0.01–0.20 having slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement and 0.81–0.99 high/almost perfect agreement (McHugh, 2012). Differences between the groups were assessed with the chi-square test with an *a-priori* alpha level of less than 0.05. Frequency of deviations for ‘novice’ and ‘experienced’ raters were also reported. STATA (Version 16, StataCorp, College Station, TX), was used to perform the data analysis.

**Post-hoc optimization of reliability**—If reliability of the LSDT did not yield kappa scores greater than 0.80 (McHugh, 2012), then additional analysis would be performed using a 2 category movement appraisal system (Rabin et al., 2014). The 2 category system will be analyzed similar to Rabin et al. (2014), who converted the classic 3 category system (‘good’, ‘fair’, ‘poor’) to a 2 category system of ‘good’ movement (0–1 deviations) and ‘moderate’ movement (2 deviations or more). Rabin et al. (2014) made this change because the sample of interest (Israel Defense Forces soldiers with PFP) was rarely scoring ‘poor’ on the 3 category system, and yielded high levels of agreement (91%) amongst raters and high reliability (κ = 0.81) with the LSDT.

**Results**

The overall inter-rater reliability for the LSDT for physical therapists was fair at baseline and moderate at 1 week post testing (Table 4). Separately, ‘experienced’ and ‘novice’ groups both yielded the same moderate inter-rater reliability level at baseline, and at 1 week post testing. There was no statistical difference between ‘experienced’ or ‘novice’ group ratings at baseline and 1 week post testing. Intra-rater reliability was moderate for the ‘experienced’ group and ‘novice’ group, with no statistical difference for each group from baseline to 1 week post testing. Test-retest reliability for ‘experienced’ raters was moderate.
\( \kappa = 0.57 \) with 71.2\% agreement in scores from baseline to 1 week and no statistically significant difference \((p = .55)\). Results were similar for ‘novice’ raters with moderate test-retest reliability \((\kappa = 0.45)\) with 65.1\% agreement in scores from baseline to 1 week and no statistically significant difference \((p = .73)\). Majority of participants in the study were classified by physical therapy raters as ‘poor’ (Figure 2).

There were differences in reliability for the subcategories of movement appraisal for the LSDT (‘good’, ‘fair’, ‘poor’) (Table 4). For ‘good’ and ‘poor’ quality of movement on the LSDT, ‘experienced’ and ‘novices’ yielded moderate to substantial reliability. ‘Fair’ quality of movement on the LSDT yielded slight agreement at baseline and fair agreement at 1 week post testing for ‘experienced’ and ‘novices’. Overall percent agreement between all raters was 60.6\% at baseline and 77.3\% at 1 week post testing, with each rater percent agreement reported in Table 5.

For post-hoc optimization of reliability calculations, a 2 category system was analyzed where we converted the classic 3 category system (‘good’, ‘fair’, ‘poor’) to a 2 category system of ‘good’ movement (0–1 deviations) and ‘moderate’ movement (2 deviations or more). Moderate reliability was yielded with the 2 category system at baseline \((\kappa = 0.56)\) and 1 week \((\kappa = 0.58)\).

**Discussion**

Inter- and intra-rater reliability of the LSDT using the 3 category method ranged from fair to moderate. There was no difference in reliability between ‘experienced’ and ‘novice’ physical therapists at baseline or 1 week post testing. The overall inter-rater reliability for the LSDT for both ‘experienced’ and ‘novice’ physical therapists together was greater at 1 week post testing as compared to baseline. The improved reliability may suggest the physical therapist raters became more familiar with the testing process, more familiar with the LSDT, and seeing the videos another time. Only moderate reliability was yielded with the 2 category system. Our study adds to our understanding of the LSDT as we have demonstrated that there is no difference between ‘experienced’ and ‘novice’ physical therapists when using the test in participants with chronic PFP.

Raters had greater disagreement when evaluating the movement quality of participants that were not distinctly ‘good’ or ‘poor’. This is reflected in Table 4 with raters yielding only slight agreement when it came to identifying participants with ‘fair’ quality of movement. This gives more credence to the idea that broader grouping within a classification tool, could lead to improved reliability. This was an issue identified in another study assessing the reliability of the LSDT (Chmielewski et al., 2007). Rabin et al. (2014) used a dichotomous system of ‘good’ movement quality (0–1 deviations) or ‘moderate’ (≥2 deviations) with the LSDT in a group of Israel Defense Forces soldiers referred to physical therapy for PFP and yielded 91\% agreement and good reliability \((\kappa = 0.81)\) between 2 raters. We only yielded moderate reliability with a 2 category system, which may be explained by our study participants being more heterogeneous, differences in chronicity of PFP (>3 months vs. >4 weeks), and majority of participants being classified as ‘poor’ with the 3 category system (Figure 2). Rabin et al. (2014) elected to use a 2 category system since the sample of Israel
Defense Forces soldiers with PFP was rarely scoring ‘poor’ on the 3 category system. Others have found good reliability when using a 2 category appraisal system (i.e., “truly high-risk” vs. “truly low risk”) during the drop jump task when evaluating dynamic knee valgus (Ekegren et al., 2009), and improved reliability with broader classifications when evaluating scapular dyskinesia (Kibler et al., 2002; McClure et al., 2009; O’Connor, McCaffrey, Whyte, and Moran, 2016).

Comparison of findings with those reported in the literature

Other studies have investigated the LSDT (Chmielewski et al., 2007; Piva et al., 2006; Rabin and Kozol, 2010; Rabin et al., 2014), however this is the first study to investigate differences between ‘experienced’ and ‘novice’ raters in participants with chronic PFP (Table 6). Only 2 previous studies investigated the reliability in a population with PFP and 2 other studies used a healthy population (Chmielewski et al., 2007; Piva et al., 2006; Rabin and Kozol, 2010; Rabin et al., 2014). Our reliability scores at 1 week post are similar to Piva et al. (2006) as both studies have a similar age and sex distribution of participants, with main differences being the chronicity of PFP and number of raters. Having more raters allowed us to investigate differences between ‘experienced’ and ‘novice’ physical therapists. The greater number of raters could be a potential reason why the raters yielded a lower kappa score at baseline testing. Rabin et al. (2014) accomplished almost perfect reliability and high agreement in a group of soldiers that were significantly younger, and likely more active than the sample in our study. The step height was lower for Rabin et al. (2014) at 15 cm, whereas our step height was standardized at 20 cm (Nakagawa, Serrão, Maciel, and Powers, 2013). This could lead to more movement deviations, with participants having to reach farther to tap their heel to the ground. Also, the majority of our sample was classified as having ‘poor’ movement quality (Figure 2), whereas the sample in Rabin et al. (2014) had better movement quality. The entire sample studied by Rabin et al. (2014) were soldiers, which suggests the sample likely participated in similar activities and exercise regimen, whereas we did not control for this in our sample of participants. In summary, it appears the participants in Rabin et al. (2014) were younger and had better movement quality than the population we studied.

Two other studies investigated the reliability in a healthy sample of participants and yielded low (Chmielewski et al., 2007) to moderate (Rabin and Kozol, 2010) reliability scores. Of these studies analyzing a healthy population, the study with 100% females from an undergraduate university yielded the higher reliability score (Rabin and Kozol, 2010) compared to a sample of participants of similar age where sex was not controlled (Chmielewski et al., 2007). This may suggest that the sample needs to be homogenous even in a healthy sample of participants without PFP to yield moderate to high reliability scores. It is important to note that the Chmielewski et al. (2007) study did not assess knee alignment similarly to what previous studies did (Piva et al., 2006; Rabin and Kozol, 2010; Rabin et al., 2014).

Limitations

Participants with PFP were not rated in real time by our raters. This caused a concern for external validity of the study. A face-to-face assessment of the LSDT would be more
relevant for clinical practice than video assessment. To account for this, raters were unable to pause the videos of the LSDT while scoring to mimic the real time flow of rating a participant in clinic. The same identifying numbers were used for each participant with PFP and the videos were presented in the same order from each testing session, which could increase the chance of rater “memory” altering the score. Although not specifically asked, it is unlikely a rater could remember the scores they selected at baseline because of the second scoring session taking place a week apart and the inability for the rater to pause the video while a participant performed the LSDT.

Although it is a strength of the study to have 2 GoPro cameras capture movement in both the frontal and sagittal planes of movement, the camera capturing the lateral view can only help capture 1 criterion of the LSDT and may offer little benefit compared to the other camera view in front of the participant. Box height was not adjusted based on participants height and could have impacted our results. We standardized our box height at 20 cm similar to previous studies (Silva et al., 2019). Future studies should adjust step height based on participant height (Chmielewski et al., 2007), or rule out step height as a contributing factor by measuring ipsilateral knee flexion angle when contralateral heel contacts ground (Rabin et al., 2014). For the physical therapist raters, years of physical therapy experience and post-professional training are not an indication that the rater had specific experience with the LSDT, or the LSDT with 2D video analysis. Therefore, years of clinical experience does not normalize the number of step down assessments performed by a physical therapist. Even though reliability at 1 week post was better than baseline reliability, suggesting more experience with the LSDT improved reliability, there was no statistical difference between reliability at baseline and 1 week post.

Conclusions

Using categories (‘good’, ‘fair’, ‘poor’) to appraise movement quality during the LSDT for a general population with chronic unilateral PFP is limited due to moderate reliability. Clinical experience and board specialty certifications do not influence the score on the LSDT. Future research should assess for reliability of the LSDT using a simple 2 categorization system on different homogenous groups of participants with the goal of improving reliability and rater agreement.

Acknowledgments

The authorship team would like to thank Charlie Domnisch, Kayla Borchers, Evan Luse, and Greg Hock for helping with the study as physical therapist raters. The project described was supported by Award Number UL1TR001070 from the National Center For Advancing Translational Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center For Advancing Translational Sciences or the National Institutes of Health. Also, support was received from a grant from the Ohio Physical Therapy Association.

Funding

This work was supported by the The Ohio Physical Therapy Association; The Ohio State University Center for Clinical and Translational Science [UL1TR002733].
References

American Board of Physical Therapy Specialties 2017 Minimum eligibility requirements and general information for all physical therapist specialist certification examinations. American Physical Therapy Association. https://www.apta.org/your-career/career-advancement/specialist-certification.

Chmielewski TL, Hodges MJ, Horodyski M, Bishop MD, Conrad BP, Tillman SM 2007 Investigation of clinician agreement in evaluating movement quality during unilateral lower extremity functional tasks: A comparison of 2 rating methods. Journal of Orthopaedic and Sports Physical Therapy 37: 122–129. 10.2519/jospt.2007.2457. [PubMed: 17416127]

Ekegren CL, Miller WC, Celebrini RG, Eng JJ, Macintyre DL 2009 Reliability and validity of observational risk screening in evaluating dynamic knee valgus. Journal of Orthopaedic and Sports Physical Therapy 39: 665–674. 10.2519/jospt.2009.3004. [PubMed: 19721212]

Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O’Neal L, McLeod L, Delacqua G, Delacqua F, Kirby J, et al. 2019 The REDCap consortium: Building an international community of software platform partners. Journal of Biomedical Informatics 95: 103–208. 10.1016/j.jbi.2019.103208.

Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG 2009 Research electronic data capture (REDCap) - A metadata-driven methodology and workflow process for providing translational research informatics support. Journal of Biomedical Informatics 42: 377–381. 10.1016/j.jbi.2008.08.010. [PubMed: 18929686]

Jones D, Tillman SM, Toft K, Mizner RL, Greenberg S, Moser MW, Chmielewski TL 2014 Observational ratings of frontal plane knee position are related to the frontal plane projection angle but not the knee abduction angle during a step-down task. Journal of Orthopaedic and Sports Physical Therapy 44: 973–978. 10.2519/jospt.2014.4960. [PubMed: 25366084]

Kibler WB, Uhl TL, Maddux JW, Brooks PV, Zeller B, McMullen J 2002 Qualitative clinical evaluation of scapular dysfunction: A reliability study. Journal of Shoulder and Elbow Surgery 11: 550–556. 10.1067/mse.2002.126766. [PubMed: 12469078]

McClure P, Tate AR, Kareha S, Irwin D, Zlupko E 2009 A clinical method for identifying scapular dyskinesis, part 1: Reliability. Journal of Athletic Training 44: 160–164. 10.4085/1062-6050-44.2.160. [PubMed: 19295960]

McGovern RP, Martin RL, Christoforetti JJ, Kivlan BR 2018 Evidence-based procedures for performing the single leg squat and step-down tests in evaluation of non-arthritic hip pain: A literature review. International Journal of Sports Physical Therapy 13: 526–536. 10.26603/ijspt20180526. [PubMed: 30038839]

McHugh ML 2012 Interrater reliability: The kappa statistic. Biochemia Medica 22: 276–282. 10.11613/BM.2012.031. [PubMed: 23092060]

Nakagawa TH, Maciel CD, Serrão FV 2015 Trunk biomechanics and its association with hip and knee kinematics in patients with and without patellofemoral pain. Manual Therapy 20: 189–193. 10.1016/j.math.2014.08.013. [PubMed: 25261089]

Nakagawa TH, Serrão FV, Maciel CD, Powers CM 2013 Hip and knee kinematics are associated with pain and self-reported functional status in males and females with patellofemoral pain. International Journal of Sports Medicine 34: 997–1002. 10.1055/s-0033-1334966. [PubMed: 23771827]

O’Connor S, McCaffrey N, Whyte E, Moran K 2016 The development and reliability of a simple field-based screening tool to assess for scapular dyskinesis. Journal of Sports Rehabilitation 20: 40–44.

Piva SR, Fitzgerald GK, Irrgang JJ, Fritz JM, Wisniewski S, McGinty GT, Childs JD, Domenech MA, Jones S, Delitto A 2009 Associates of physical function and pain in patients with patellofemoral pain syndrome. Archives of Physical Medicine and Rehabilitation 90: 285–295. 10.1016/j.apmr.2008.08.214. [PubMed: 19236982]

Piva SR, Fitzgerald K, Irrgang JJ, Jones S, Hando BR, Browder DA, Childs JD 2006 Reliability of measures of impairments associated with patellofemoral pain syndrome. BMC Musculoskeletal Disorders 7: 33. 10.1186/1471-2474-7-33. [PubMed: 16579850]

Rabin A, Kozol Z 2010 Measures of range of motion and strength among healthy women with differing quality of lower extremity movement during the lateral step-down test. Journal of Physiother Theory Pract. Author manuscript; available in PMC 2023 December 01.
Rabin A, Kozol Z, Moran U, Efergan A, Geffen Y, Finestone AS 2014 Factors associated with visually assessed quality of movement during a lateral step-down test among individuals with patellofemoral pain. Journal of Orthopaedic and Sports Physical Therapy 44: 937–946. 10.2519/jospt.2014.5507. [PubMed: 25347229]

Rabin A, Portnoy S, Kozol Z 2016 the association of ankle dorsiflexion range of motion with hip and knee kinematics during the lateral step-down test. Journal of Orthopaedic and Sports Physical Therapy 46: 1002–1009. 10.2519/jospt.2016.6621. [PubMed: 27686412]

Rodeghero J, Wang YC, Flynn T, Cleland JA, Wainner RS, Whitman JM 2015 The impact of physical therapy residency or fellowship education on clinical outcomes for patients with musculoskeletal conditions. Journal of Orthopaedic and Sports Physical Therapy 45: 86–96. 10.2519/jospt.2015.5255. [PubMed: 25579690]

Silva RL, Pinheiro YT, De Almeida Lins CA, De Oliveira RR, Silva RS 2019 Assessment of quality of movement during a lateral step-down test: Narrative review. Journal of Bodywork and Movement Therapies 23: 835–843. 10.1016/j.jbmt.2019.05.012. [PubMed: 31733769]

Willy RW, Hoglund LT, Barton CJ, Bolgla LA, Scalzitti DA, Logerstedt DS, Lynch AD, Snyder-Mackler L, McDonough CM, Altman R et al. 2019 Patellofemoral pain. Journal of Orthopaedic and Sports Physical Therapy 49: CPG1–CPG95. 10.2519/jospt.2019.0302.
Figure 1.
Lateral step down test.
Figure 2.
Frequency of classifications at baseline and 1 week.
Table 1.

Lateral step down test scoring criteria.

| Movement Deviation | Interpretation                          | Score |
|--------------------|-----------------------------------------|-------|
| Arm Strategy       | Removal of a hand from the wait         | 1     |
| Trunk Alignment    | Leaning in any direction                | 1     |
| Pelvic Plane       | Loss of horizontal plane                | 1     |
| Knee Position      | Tibial tuberosity medial to second toe  | 1     |
|                    | Tibial tuberosity medial to medial border of foot | 2     |
| Steady Stance      | Stepping down on contralateral limb or foot wavering | 1     |

Cumulative Score

0 or 1 Good
2 or 3 Fair
4 or above Poor
### Table 2.

Rater Characteristics.

|                          | ‘Experienced’ Raters (n = 3) | ‘Novice’ Raters (n = 3) |
|--------------------------|-----------------------------|------------------------|
| Age [mean (range)]       | 32.3 years old (29–39)      | 26.3 years old (26–27) |
| Experience [mean (range)]| 8.3 years (5–14)            | 1.7 years (1–2)        |
| DPT (%)                  | 100%                        | 100%                   |
| AT (%)                   | 66.7%                       | 0%                     |
| Board Certification (%)  | 100% (2 SCS, 1 OCS)         | 0%                     |

DPT: Doctor of Physical Therapy; AT: Athletic Trainer; SCS: Board Certified in Sports Physical Therapy; OCS: Board Certified in Orthopedic Physical Therapy.
### Table 3.

Participant characteristics.

| Variable     | n  | Mean (SD) | Range    |
|--------------|----|-----------|----------|
| Age (years old) | 22 | 27.8 (6.6) | 18 to 38 |
| Sex          | 22 | 0.40 (0.50) | 0 to 1   |
| BMI (kg/m\(^2\)) | 21* | 26.4 (4.7) | 19.8 to 37.3 |

BMI: Body Mass Index

*Height and weight data was missing for 1 female
Table 4.
Differences in reliability between experienced and novice physical therapists.

| Measurement          | Overall Agreement between ‘Experienced’ and ‘Novice’ Raters (Kappa (percent agreement)) | ‘Experienced’ Raters (n = 3) Weight average Kappa | ‘Novice’ Raters (n = 3) Weight average Kappa | Difference Between Groups p-value |
|----------------------|---------------------------------------------------------------------------------------|-----------------------------------------------|---------------------------------------------|----------------------------------|
| Lateral Step Down    | 0.40 (60.6%)                                                                            | Outcome Good 0.73 Kappa                       | Outcome Good 0.42 Kappa                    | 0.13                             |
| Test                 |                                                                                       | Fair 0.07 Kappa                              | Fair 0.11 Kappa                             |                                  |
| Baseline Score       |                                                                                       | Poor 0.41 Kappa                              | Poor 0.64 Kappa                             |                                  |
| Combined             |                                                                                       | 0.41 Combined Kappa                          | 0.41 Combined Kappa                        |                                  |
| Lateral Step Down    | 0.65 (77.3%)                                                                            | Outcome Good 0.57 Kappa                       | Outcome Good 0.46 Kappa                    | 0.94                             |
| Test                 |                                                                                       | Fair 0.36 Kappa                              | Fair 0.28 Kappa                             |                                  |
| 1 Week Post Score    |                                                                                       | Poor 0.68 Kappa                              | Poor 0.69 Kappa                             |                                  |
| Combined             |                                                                                       | 0.53 Combined Kappa                          | 0.48 Combined Kappa                        |                                  |
Table 5.

Kappa and percent agreement among raters at baseline.

|           | Experienced 1 | Experienced 2 | Experienced 3 | Novice 1 | Novice 2 | Novice 3 |
|-----------|---------------|---------------|---------------|----------|----------|----------|
| Experienced 1 | 0.31 (54.5%) | 0.48 (63.6%) | 0.64 (77.3%) | 0.61 (77.3%) | 0.46 (63.6%) |        |
| Experienced 2 | 0.31 (54.5%) | 0.46 (63.6%) | 0.26 (50%)   | 0.38 (59.1%) | 0.21 (45.4%) |        |
| Experienced 3 | 0.48 (63.6%) | 0.46 (63.6%) | 0.39 (59.1%) | 0.30 (50.0%) | 0.15 (45.4%) |        |
| Novice 1     | 0.64 (77.3%) | 0.26 (50%)   | 0.39 (59.1%) | 0.47 (68.2%) | 0.35 (59.1%) |        |
| Novice 2     | 0.61 (77.3%) | 0.38 (59.1%) | 0.30 (50.0%) | 0.47 (68.2%) | 0.45 (62.6%) |        |
| Novice 3     | 0.46 (63.6%) | 0.21 (45.4%) | 0.15 (45.4%) | 0.35 (59.1%) | 0.45 (62.6%) |        |
Table 6.

Reliability of the lateral step down test in the literature.

| Author & Year       | n   | Characteristics and chronicity                                                                 | Sex (%) female | Age (mean ± SD yrs) | Step Height | # of Repetitions | # of Raters | # categories | Kappa/ % Agreement |
|---------------------|-----|-----------------------------------------------------------------------------------------------|----------------|---------------------|-------------|------------------|-------------|--------------|-------------------|
| Piva et al. 2006    | 30  | Diagnosed with PFP by physician with duration for at least 4 weeks                             | 56.7%          | 29.1±8.4            | 20 cm       | 5                | 4           | 3            | κ=0.67 (80%)       |
| Rabin et al. 2014   | 79  | Israel Defense Forces soldiers with PFP by physician with duration for at least 4 weeks       | 50.6%          | 19.1±1.5            | 15 cm       | 6 (last 5 assessed only) | 2           | 2            | κ= 0.81 (91%)      |
| Data from this study| 22  | Participants with PFP for at least 3 months                                                   | 63.6%          | 28.2 ±6.5           | 20 cm       | 5                | 6           | 3            | κ= 0.40-0.65 (60.6%-77.3%) |

| Author & Year       | n   | Characteristics                                                                                     | Sex (%) female | Age (mean ± SD) | Step Height | # of Repetitions | # of Raters | # categories | Kappa/ % Agreement |
|---------------------|-----|---------------------------------------------------------------------------------------------------|----------------|---------------|-------------|------------------|-------------|--------------|-------------------|
| Chmielewski et al. 2007 | 25 | Healthy participants with no low back or lower extremity pain                                      | 72%            | 22.4±1.3      | 15.24 cm or 25.4 cm depending on height of participant | ?            | 3           | 3/4            | κ= 0.19 (55-82%)    |
| Rabin & Kozol 2010  | 29  | Undergraduate female students from a university campus                                             | 100%           | 24.3±3.2      | Adjusted so participant can maintain 60° of knee flexion upon heel contact | 5            | 2           | 3             | κ= 0.59 (83%)       |

SD: standard deviation; PFP: patellofemoral pain