Inter-rater and intra-rater reliability of the Triple Joint Flexion Test (TJFT) scores in adolescent field and court athletes by novice raters
Declaration

Name of candidate: Darryl Keith Jenkins

This thesis entitled “Inter-rater and intra-rater reliability of the Triple Joint Flexion Test (TJFT) scores in adolescent field and court athletes by novice raters” is submitted in partial fulfilment for the requirements for the Unitec degree of Master of Osteopathy

Candidate’s declaration:

I confirm that:

- This thesis represents my own work;
- Research for this work has been conducted in accordance with the Unitec Research Ethics Committee Policy and Procedures, and has fulfilled any requirements set for this project by the Unitec Research Ethics Committee.

Research Ethics Committee Approval Number: UREC 2013-1019

Candidate Signature: .......................... Date: ..................

Student number: 1381338
Abstract

Background  Despite the common use of lower limb movement screening tools, the limitations for clinical use are not well understood. In a movement screening context, there are two sources of variation associated with reliability measurements. These are: (1) the athlete's performance of the test; and (2) the observational ratings of the test movements made by the rater. The Triple Joint Flexion Test (TJFT) is a new movement screening tool based on work by Tainhague, (2015) with development level adolescent court and field athletes. To date, one previous study has reported objective data of intra-athlete within- and between-session variability of the Triple-Joint Flexion Test (TJFT). To date, no objective data of observational ratings of the TJFT sub-test scores has been reported. This thesis is divided into 3 sections: (1) A review of the literature (2) A study of the inter- and intra-rater reliability of the Triple Joint Flexion Test reported in the form of a journal manuscript; and (3) supplementary material.

Design  Video based, repeated measures

Aims  To establish the inter- and intra-rater reliability of the Triple Joint Flexion Test on screening adolescent male court and field athletes.

Method  Seven trained novice raters rated TJFT sub tests on two occasions using real time video. The TJFT was performed by 17 adolescent court and field athletes and 182 ratings were made on each occasion.

Results  Inter-rater weighted agreement of the six TJFT sub tests by 7 raters on Day 1 and 6 raters on Day 2 demonstrated ‘moderate’ (AC2 > 0.41) to ‘substantial’ inter-rater agreement (AC2 > 0.61). Intra-rater weighted agreement of the six TJFT sub-tests by 6 raters was mostly ‘moderate’ (AC2 > 0.41) and ranged from ‘fair’ (AC2 > 0.21) to ‘near perfect’ (AC2 > 0.81).

Conclusion  The inter-rater and intra-rater agreement of TJFT sub-test scores have now been investigated, and found to sufficiently reliable. It is un-common for both athlete and rater sources of variability to be identified for a lower limb movement screen. As such, clinicians may use the TJFT with a greater understanding of the expected variability from both the rater and the athlete. In the wider movement screening literature, reports of rater-reliability for trained novices scoring multisegmental movements in real time is unique. This is also the first scoring protocol to have reported acceptable levels of inter and intra-rater agreement for a higher threshold movement task under these scoring conditions

Keywords  Movement screening, Reliability, Injury, Lower extremity, Triple Joint Flexion Test, Pre Participation evaluation.
I would like to wholeheartedly thank everyone who supported me over the last six years of my academic career. This would have been a lot harder without all your help.

This journey would have never started without the encouragement and support from my incredible parents, who continue to support my (albeit questionable) life endeavours. Thank you both so much. I’m also very lucky to have a large family and a close group of friends who never did mind me disappearing for months at a time, and were always there when I needed a laugh.

I would have never reached this point in my studies if it wasn’t for my cousin Matthew, and my friend Rosalie who both patiently put up with and corrected my sub-optimal academic writing skills, time after time after time. Thank you both 😊

This research project was gifted to me by Matiu Taingahue, who I would like to thank for opening my eyes to a piece of manual therapy I have become completely engrossed in.

This research project would have never been completed if not for Rob Moran, my research supervisor. Thank you Rob for teaching me to take a step back and look at the whole picture and to be critical of what I was seeing.

Most of all, I would like to thank my partner Jessica Rioufrays. This would have been a very dull 5 year journey without you and I’m lucky to have someone as courageous, smart and beautiful as you by my side.
# Table of Contents

ABSTRACT ........................................................................................................................................ III

ACKNOWLEDGEMENTS ...................................................................................................................... IV

TABLE OF CONTENTS ......................................................................................................................... V

SECTION 1: INTRODUCTION TO THESIS ........................................................................................ - 1 -

1.1 HEALTH BENEFITS AND POTENTIAL HARMs OF PHYSICAL ACTIVITY AND SPORT .......... - 1 -

1.2 MITIGATING INJURIES AS A RESULT OF PHYSICAL ACTIVITY AND SPORT ....................... - 1 -

1.3 THE TRIPLE JOINT FLEXION TEXT ...................................................................................... - 3 -

2 LITERATURE REVIEW ................................................................................................................... 5

2.1 RISK AND BENEFITS OF PHYSICAL ACTIVITY AND SPORTS .............................................. 5

2.2 HOW ARE INJURIES MITIGATED IN SPORT AND PHYSICAL ACTIVITY? ............................. 6

2.2.1 Pre-participation evaluation ................................................................................................. 6

2.2.2 The concept of movement screening .................................................................................... 7

2.2.3 A rationale for lower extremity movement screens ............................................................... 7

2.2.4 What lower extremity movement screens exist and are they reliable? ............................... 8

2.2.5 Example of movement screens originating within the literature ....................................... 9

2.2.6 Example of movement screens originating from clinical practice ..................................... 10

2.2.7 Why does reliability matter? ............................................................................................. 11

2.3 WHAT SOURCES OF VARIABILITY WITHIN A RATING ARE THERE? ............................... 12

2.3.1 Sources of variability attributable to athlete performance ............................................... 12

2.3.2 Sources of variability attributable to the rater ........................................................................ 13

2.3.3 Summary of rater source of variability ................................................................................ 16

2.4 A METHODOLOGICAL CRITIQUE OF EXISTING RELIABILITY STUDIES ....................... 18

2.4.1 Introduction .......................................................................................................................... 18

2.4.2 Method of review .................................................................................................................. 18

2.4.3 Findings .................................................................................................................................. 19

2.4.4 Summary ................................................................................................................................ 24

2.5 OVERALL SUMMARY OF THE EXISTING LITERATURE TO DESCRIBE THE RATIONALE FOR MOVEMENT SCREENING AND IDENTIFY THE AREA OF RESEARCH NEED ......................................................... 26

REFERENCES ...................................................................................................................................... 29

....................................................................................................................................................... 36
SECTION 2: MANUSCRIPT .................................................................................................................. 1

1 ABSTRACT ......................................................................................................................................... 3

BACKGROUND ....................................................................................................................................... 3
RESULTS .............................................................................................................................................. 3
KEYWORDS ......................................................................................................................................... 3

2 INTRODUCTION ................................................................................................................................ 4

3 METHODS ......................................................................................................................................... 5

3.1 DESIGN .......................................................................................................................................... 5
3.2 PARTICIPANTS ................................................................................................................................. 5
  3.2.1 Recruitment and description of athletes ................................................................................... 5
  3.2.2 Recruitment and description of raters ...................................................................................... 6
3.3 TJFT PROTOCOL ............................................................................................................................. 6
  3.3.1 Visual assessment procedures ................................................................................................. 6
3.4 DATA ANALYSIS ............................................................................................................................ 7

4 RESULTS .......................................................................................................................................... 8

4.1 INTER-RATER RELIABILITY ........................................................................................................... 8
4.2 INTRA-RATER RELIABILITY BETWEEN DAYS ............................................................................ 8
  4.2.1 Frontal view ........................................................................................................................... 8
  4.2.2 Sagittal view .......................................................................................................................... 8
4.3 INFLUENCE OF WEIGHTINGS ...................................................................................................... 8

5 DISCUSSION .................................................................................................................................... 12

5.1 INTRODUCTION .............................................................................................................................. 12
5.2 SCORING PROTOCOL .................................................................................................................... 12
5.3 HIGH-VELOCITY MOVEMENT SUB-TESTS .................................................................................... 13
5.4 RATER TRAINING ............................................................................................................................ 14
5.5 RATING PROCEDURE ..................................................................................................................... 14
5.6 LIMITATIONS ................................................................................................................................. 14
5.7 IMPLICATIONS FOR FUTURE RESEARCH .................................................................................... 15

6 CONCLUSION ................................................................................................................................... 16

7 REFERENCES ..................................................................................................................................... 18

8 SECTION 3 APPENDICES .................................................................................................................. 1

APPENDIX 1: DESCRIPTION OF TJFT ................................................................................................. 2
APPENDIX 2: TJFT STANDARDISED SCORE CARDS................................................................. 3

APPENDIX 3: ETHICS DOCUMENTS: CONSENT FORM, INFORMATION SHEET, APPROVAL LETTER.................. 3
Section 1: Introduction to thesis

1.1 Health benefits and potential harms of physical activity and sport
Regular physical activity contributes to the primary and secondary prevention of several chronic diseases, including cardiovascular disease (Bouchard, Shephard, & Stephens, 1994; Taylor et al., 2004) diabetes mellitus (Gibbs et al., 2014), cancer (colon and breast) (Mishra et al., 2012), obesity (Blair & Brodney, 1999), bone and joint diseases (osteoporosis and osteoarthritis) (Puett & Griffin, 1994), and depression (Ya, Jin, Oh, & Choi, 2015). Although the most physically active people are at the lowest risk of developing chronic disease, substantial improvements in health are also seen when inactive people become physically active (Ekelund et al., 2015; Hu et al., 2004). Aside from the prevention of chronic disease, there are many other reasons why people participate in sports and physical activity, including pleasure, socialisation, and competition (Dalziel, 2011). In New Zealand, an increase in government funding has led to agencies such as Sport New Zealand (formerly SPARC) to increase physical activity and sporting participation rates in New Zealand (Sport New Zealand, 2015). However, as a consequence of increasing physically active lifestyles, an increase in the number of sport and physical activity related injuries is expected (Verhagen & van Mechelen, 2010). Of injuries associated with participation in sport and physical activity, a large proportion occur in the lower extremity (Fong, Hong, Chan, Yung, & Chan, 2012; Hussain, 2010; LaBella, 2007) and often lead to short and long term negative effects (Verhagen & van Mechelen, 2010), including early onset osteoarthritis (Lohmander, Englund, Dahl, & Roos, 2007) reduced balance (Relph, Herrington, & Tyson, 2014; Parus, Lisiński, & Huber, 2015), and recurrent injury (Dunne et al, 2004; Hoch, Staton, McKeon, Mattacola, & McKeon, 2012).

1.2 Mitigating injuries as a result of physical activity and sport
Although it is generally accepted that acute impact type injuries are largely unavoidable, overuse type injuries may be avoidable where potentially modifiable injury risk factors are controlled (Verhagen & van Mechelen, 2010). Neuromuscular control is a modifiable risk factor which has achieved considerable attention in recent research (Brent, 2011; Comerford & Mottram, 2001; Lee & Powers, 2014; Mauntel et al., 2013; Mottram & Comerford, 2008; Noyes, Barber-Westin, Fleckenstein, Walsh, & West, 2005; Zazulak, Hewett, Reeves, Goldberg, & Cholewicki, 2007). Neuromuscular control can be described by the presence or absence of dynamic joint stabilization during coordinated movement and is achieved via a combination of active muscle forces and passive ligamentous restraints (Williams et al.,
Several terms are used interchangeably by academic and non-academic authors to describe neuromuscular control, and this has led to some confusion in the literature. Commonly employed terms include ‘movement quality’ (Moran, Schneiders, Major, & Sullivan, 2015) ‘movement dysfunction’ (O’Sullivan, 2000), and ‘movement impairment’ (Mischiati et al., 2015). For the purposes of this review, ‘movement quality’ will be used to describe neuromuscular control in general terms and ‘movement dysfunction/impairment’ will be used to specifically describe poor movement quality. For example, jump landing tasks are often used to assess movement quality (Ekegren et al., 2009; Padua et al., 2009). Within a jump landing movement quality test, movement dysfunction may be observed in the form of knee valgum (Ageberg et al., 2010; Ekegren, et al., 2009; Padua et al., 2009; Whatman, Hing, & Hume, 2012) or movement asymmetry (Kinzey & Armstrong, 1998; Padua et al., 2009)

There is a growing body of evidence linking movement dysfunction with increased risk of the lower extremity overuse injuries (Chuter & Janse de Jonge, 2012; Felson et al., 2013; Lee & Powers, 2014; McKeon & Hertel, 2008; Stearns, Keim, & Powers, 2013). ‘Movement Screening’ has been proposed as part of physical examination for assessment of movement quality and identifying movement dysfunction primarily through visual analysis, often using several different tests grouped together – sometimes referred to as a ‘test battery’ (Frohm, Heijne, Kowalski, Svensson, & Myklebust, 2012). Several movement screening batteries have been described in recent literature, such as the ‘Functional Movement Screen’, (Cook, Burton, & Hoogenboom, 2006a, 2006b) the ‘Athletic Ability Assessment’ (McKeown, Taylor-McKeown, Woods, & Ball, 2014) the Movement Competence Screen (Kritz, 2012) and the ‘Netball Movement Screening Tool’ (Reid, Vanweerd, Larmer, & Kingstone, 2014). As lower extremity injuries are common and potentially avoidable, several lower extremity specific movement screen have also been described, including the ‘Landing Error Scoring System’ (Padua et al., 2009) ‘Star Excursion Balance Test’ (Kinzey & Armstrong, 1998), ‘Y-balance test’ (Plisky et al., 2009), and most recently, the ‘Triple Joint Flexion Test’ (Taingahue, 2015)

There appears to be a general lack of rigorous testing in lower limb movement screening literature, and as a consequence the limitations for clinical use are not well understood (Kivlan & Martin, 2012). In order for clinicians and researchers to better understand the
limitations of a movement screen, the validity and reliability of the tests needs to be defined (Davidson & Keating, 2014). In general terms, test validity defines inferences that can be made based on test scores, and reliability is described by the repeatability of the test measurements (Davidson & Keating, 2014). In a movement screening context, there are two major sources of variation associated with reliability measurements, firstly, the athlete’s performance of the test movements; and secondly, in the ratings of the test movements made by the rater.

1.3 The Triple Joint Flexion Test
The TJFT is a relatively new movement screening tool which originated as part of injury reduction strategies for adolescent Norwegian and Qatar football (soccer) players. The test was designed to address the clinical need for identifying injury risk and to help guide physical training interventions for large sport teams and training squads. Despite the origins and clinical application of the TJFT, only one investigation of the test has been conducted (Taingahue, 2015).

It has been suggested that the two sources of rating variability within the TJFT could be explored separately in a two-part study (Taingahue, 2015). This approach would be similar to previous lower limb movement screening investigations by Whatman, Hing and Hume, (2011), and Whatman, Hing and Hume, (2012). Taingahue, (2015) has since conducted the initial investigation of the TJFT exploring the athlete variability within and between sessions. In sequence to Taingahue (2015) study, the aim of the present study was to conduct an investigation of the inter- and intra rater reliability of the TJFT.

The TJFT consists of three lower extremity movements selected to challenge mobility and neuromuscular control. The three fundamental lower limb movements (Squat, single leg squat, single leg hop and stick) are kinematically similar, but the loading characteristics of each movement progressively increase demand for dynamic stability (Taingahue, 2015). The rating criteria for each movement test are similar, and were designed to meet multiple clinical needs: (1) to permit the comparison of right/left differences; (2) to provide clinically useful information for guiding intervention; (3) to maximise the potential for reliable results; (4) to minimise the requirement for expensive or technical equipment; and (5) to simplify screening large groups of athletes as typically occurs in pre-season screening situations.

This thesis is divided into two sections. As there is no other research on the reliability of the TJFT, the first section of the thesis is a literature review of other lower limb movement screening tests. This review covers the rater reliability of the screens, the rating methods
which influence rater reliability, and the methodological quality of the studies. Following the literature review, an investigation of the TJFT rater reliability is presented in the style of a journal manuscript.

The specific aim of the study was to explore the inter- and intra-rater reliability of the TJFT. The test was performed by a group of development-level, adolescent, court and field sport athletes and scored by postgraduate osteopath students. The athlete group was the same as that used previously by Taingahue, (2015) to investigate athlete variability. To exclusively investigate rater reliability, video ratings were used to control the intra-athlete variability and specifically investigate the intra- and inter-rater reliability of movement screening tests. This approach is common in other reliability studies (McKeown et al., 2014; Ness, Taylor, Haberl, Reuteman, & Borgert, 2015; Park, Cynn, & Choung, 2013; Smith et al., 2012; Whatman, Hing, & Hume, 2012).
2 Literature Review

The aim of this review of the literature is to provide a rationale for investigating the inter- and intra-rater reliability of the TJFT. The rationale is based on a review of: (1) The management of lower extremity injuries in sport through movement screening; (2) the limitations of existing movement screening protocols; and (3) a methodological critique of studies which examine rater reliability of lower extremity movement screens.

2.1 Risk and benefits of physical activity and sports

The benefits of regular physical activity are vast, and include the primary and secondary prevention of several chronic diseases, including cardiovascular disease (Bouchard, Shephard, & Stephens, 1994; Taylor et al., 2004) diabetes mellitus (Gibbs et al., 2014), cancer (colon and breast) (Mishra et al., 2012), obesity (Blair & Brodney, 1999), bone and joint diseases (osteoporosis and osteoarthritis) (Puett & Griffin, 1994), and depression (Ya, Jin, Oh, & Choi, 2015). The positive impact of physical activity on health is so widely supported that leading medical practitioners have advocated for the “exercise vital sign” described by Sallis (2011), to be recorded alongside other vital signs such as blood pressure (Khan et al., 2012). As substantial improvements in health are seen when people become physically active (Ekelund et al., 2015; Hu et al., 2004), an increase in government funding has led to agencies such as Sport New Zealand (former SPARC) to actively promote higher participation rates in New Zealand (Sport New Zealand, 2015). However, as a natural consequence of increasing physically active lifestyles, an increase in the number of sport and physical activity related injuries is occurring (Verhagen & van Mechelen, 2010; Verhagen et al., 2015).

The incidence of sports related injuries requiring treatment in the United States has been reported to exceed that of transport-related injuries, and although the consequences of sports related injuries were generally mild, 20 to 28% of people lost at least one day of work or school as a result (Conn, Annest, & Gilchrist, 2003). Of injuries which occur from sport or physical exercise, most occur in the lower extremity (Fong et al., 2012; Hussain, 2010; LaBella, 2007) and often lead to short and long-term negative effects (Maffuli, Longo, Gougoulias, Caine, & Denaro, 2011; Verhagen & van Mechelen, 2010) including early onset osteoarthritis (Lohmander, Englund, Dahl, & Roos, 2007), reduced balance (Relph, Herrington, & Tyson, 2014; Parus, Lisiński, & Huber, 2015), and elevated risk of recurrent injury (Dunne et al, 2004; Hoch, Staton, McKeon, Mattacola, & McKeon, 2012). Inspection
of the New Zealand Accident Compensation Corporation’s injury statistics tool (http://www.acc.co.nz/about-acc/statistics/injury-statistics-tool/index.htm#) between 2010 to 2015 shows an increase of 40,000 claims over this period for lower limb sporting injuries and associated treatment costs of NZD$140,000,000.

2.2 How are injuries mitigated in sport and physical activity?

2.2.1 Pre-participation examination

Pre-participation examination (PPE) was first described by Garrick and Requa (1978), as an assessment which enabled athletes to identify and declare health related issues for insurance or legal purposes and as a means to provide coaches with individual measures of fitness, such as blood pressure and running endurance (Garrick & Requa, 1978). In 2010, five organisations, including the American Osteopathic Academy of Sports Medicine, and the American Medical Society for Sports Medicine, joined the American College of Sports Medicine (ACSM) in publishing the 4th edition of the Pre-participation Physical Evaluation (PPE-4). The PPE-4 provides the most current comprehensive guidelines for PPE as an important component of promoting athlete health and safety in training and competition (Bernhardt, Roberts, 2010).

Although general health and medical investigations are important components of PPE (Maffey & Emery, 2006; Sanders, Blackburn, & Boucher, 2013) the physical screening components have developed immensely in recent years. PPE now includes a battery of neuromusculoskeletal tests used in the context of sport-specific biomechanical and physiological demands seen in athletic training and competition (Mottram & Comerford, 2008). In addition, several authors have suggested that there are key movements which underpin the fundamentals of athletic performance (Cook et al., 2006a; McKeown et al., 2014) and that these movements should be assessed to ensure an athlete is prepared for the wide variety of movement demands placed on them during sport. Such movements have been described as ‘real movement’ (Mottram & Comerford, 2008), ‘functional’ or ‘fundamental movement’ (Cook et al., 2006a, 2006b; Kritz, 2012) and ‘physical performance measures’ (Tarara, Hegedus, & Taylor, 2014). In the context of this review the term functional movement will be used. Examples of functional movements include squatting (Cook et al., 2006a; Taingahue, 2015; Tarara et al., 2014), hopping (McKeown et al., 2014; Taingahue, 2015), and push ups (Cook et al., 2006a; Kritz, 2012). These movements are often observed in sport and sport specific training. Collectively, the process of evaluating
functional movements in a sport specific context as a measure of neuromuscular control is referred to as a ‘movement screen’.

2.2.2 The concept of movement screening

Movement screening is a common skill practiced by sports medicine clinicians and is a key component of the clinical decision making process (Bernhardt, Bate, & Matyas, 1998; Sahrmann, 2013). Clinicians use the movement screening process to identify potentially modifiable movement quality related impairments such as, reduced ranges of motion, painful ranges of motion, balance impairments, and left to right movement asymmetry (Cook et al., 2006a; Kritz, 2012; McKeown et al., 2014; Shaffer et al., 2013; Taingahue, 2015; Tarara et al., 2014). These impairments are identified by raters scoring movements against pre-determined and standardised scoring systems which highlight movement quality related impairments (Frohm et al., 2012). The identification of movement impairments may then help guide further physical investigations and injury prevention programs based on the underlying premise that correcting these impairments will decrease the likelihood for injury and re-injury risk.

Several movement screens have been described in recent literature, and these can be generally divided into broad movement screens, and regional movement screens. Broad movement screens include movement assessment of the entire body, such as the Functional Movement Screen (FMS) (Cook et al., 2006a, 2006b), or the Athletic Ability Assessment (AAA) (McKeown et al., 2014). Regional movement screens include specific movement tests of the upper or lower body only, such as the Y-balance upper quadrant test for the upper body (Westrick, Miller, Carow, & Gerber, 2012), and the TJFT for the lower limb and pelvis (Taingahue, 2015).

2.2.3 A rationale for lower extremity movement screens

In athletic populations lower extremity injuries are common, and may lead to impaired athletic performance (Monaghan, Delahunt, & Caulfield, 2006) or the termination of sporting careers (Ristolainen, Kettunen, Kujala, & Heinonen, 2012). Injuries to the knee constitute a substantial proportion of sport related injury (Fong et al., 2012; Keogh, Hume, & Pearson, 2006). A retrospective injury study of top-level Finnish male and female athletes found that in football (soccer) athletes, knee injuries were the most common site of career ending injuries (Ristolainen et al., 2012). Knee pain is also the leading cause of reduced physical performance worldwide and is thought to precede the development of osteoarthritis
Injuries to the ankle are reported as the second most common site of athletic injury (Doherty et al., 2014; Fong et al., 2012), and ankle injuries often develop into performance diminishing chronic ankle instability in young people (Hershkovitch et al., 2015). Groin/hip injuries are also common in the athletic population, particularly in rugby union which identifies hip/groin injuries as the fourth most common site of injury (Ryan, DeBurca, & McCrereh, 2014). Although it is generally accepted that acute traumatic injuries are unavoidable, these findings across the international athletic population underscore the importance of identifying modifiable and intrinsic lower extremity injury risk factors as the first step of reducing preventable injury.

2.2.4 What lower extremity movement screens exist and are they reliable?

In recent years, a range of lower extremity movement screens have been described in the literature, some of which include a ‘battery’ of test movements, for example Whatman et al., (2012) describes a lower extremity battery consisting of 4 movements (Small knee bend single leg, lunge, hop lunge). The TJFT is another example of a test battery, and consists of 3 movements (Squat, Single Leg Squat, Single Leg Hop and Stick) (Taingahue, 2015). Other authors have described movement screens consisting of a single movement test, such as a drop jump (Boling, Thigpen, Padua, & Marshall, 2005; Ekegren, Miller, Celebrini, Eng, & Macintyre, 2009), single leg step down (Park et al., 2013; Piva et al., 2006), or single leg squat (Ness et al., 2015; Poulsen & James, 2011).

Regardless of the number of movements included in the test, lower extremity movement screens can be broadly divided into two investigative subtypes: Firstly, those screens originating in the research environment (Ageberg et al., 2010; Chmielewski et al., 2007; Ekegren et al., 2009; Whatman et al., 2012; Whatman, Hume, & Hing, 2013); and secondly, screens which originated in clinical practice (Kinzey & Armstrong, 1998; Padua et al., 2009; Taingahue, 2015). Screening batteries originating in research environments tend to investigate specific research questions around the validity and reliability of movement screening. Although these movement screens may eventually be used in clinical practice, there is a tendency for the testing procedures to be less useful to clinicians as the object of the test may not be to identify clinically meaningful dysfunction. Comparatively, screening batteries originating in clinical practice tend to have higher potential for clinical application.
as the objective of the test is to identify meaningful dysfunction, such as impaired dynamic balance (Kinzey & Armstrong, 1998), or dynamic landing errors indicative of knee injury risk (Boling et al., 2005). However, these movement screens require empirical scrutiny of reliability and validity measures to affirm their measurement properties.

2.2.5 Example of movement screens originating within the literature

The investigation by Chmielewski et al., (2007) is an example of a movement screen originating in the research environment. Chmielewski et al’s, (2007) movement screen was designed to investigate the inter and intra rater agreement when scoring functional movements by comparing two common scoring methods (overall and segmental). Whilst the movements used and the scoring methods investigated were reflective of current clinical practice, the aim of the investigation was not to investigate a movement screen for practical clinical use. Instead, the authors undertook an early step in identifying what measurement variables made for a reliable movement screen. The findings of Chmielewski et al., (2007) investigation were later used by several authors aiming to identify the components of lower extremity movement screens most likely to be scored with acceptable rater reliability (Ekegren et al., 2009; Whatman et al., 2012, Whatman et al.,2013). The common components discussed by these authors include rating criteria, scoring scales, and the velocity of the movements tested. (See figure 1).

![Movement Screen Diagram](image)

Figure 1. Commonly explored components of movement screens include (1) Movements tested (2) Scoring Scale (3) Rating Criteria. Within each component authors may choose one or more factors (dark grey box) to be included in the movement screen and this depends on the individual aim of the investigation or screening battery. As is discussed later, the Rating Criteria and Scoring Scale make up the Scoring System which ultimately guides the rating of a Movement.
2.2.6 *Example of movement screens originating from clinical practice*

As an example of a movement screen which has originated outside of the literature, the Star Excursion Balance Test (SEBT) first described by Kinzey and Armstrong, (1998) was designed to identify meaningful dynamic balance impairments in people by screening single leg squat and reach tasks in eight directions. The SEBT requires participants to squat on one leg and perform a measured reach excursion without losing balance. Kinzey and Armstrong, (1998) stated that the SEBT had been used by clinicians for some time, and that their rater reliability investigation was the first formal investigation of the test. ‘Moderate’ inter-rater agreement was reported by Kinzey and Armstrong, (1998) (ICC= 0.67 - 0.87, no confidence interval reported). Following this initial investigation, the SEBT has been reported as a reliable and useful movement screening tool by several authors across a range of populations including healthy individuals (Gribble, Kelly, Refshauge, & Hiller, 2013; Hyong & Kim, 2014; Kinzey & Armstrong, 1998), elite female football and handball players (Harøy, 2013) high school basketball players (Plisky, Rauh, Kaminski, & Underwood, 2006) and recreational athletes (Munro & Herrington, 2010). Since first described by Kinzey and Armstrong (1998), a simplified version of the SEBT, the Y-balance test (YBT) has been described (Shaffer et al. 2013; Hertel et al. 2006). The YBT consists of 3 squat and reach excursions (reduced from the original 8 directions of the SEBT) based on research indicating that these directions alone are able to identify significant reach deficits associated with impaired dynamic balance (Hertel, Braham, Hale, & Olmsted-Kramer, 2006). Good inter and intra rater reliability of the YBT has subsequently been reported (Plisky, Gorman, Butler, & Kiesel, 2009; Shaffer et al., 2013). More recently, a movement quality assessment (rather than reach distance alone) of the SEBT has been described by Ness et al., (2015) and is based on the dynamic alignment of the trunk, knee, and pelvis during the anterior squat and reach excursion. The study by Ness (2015) follows reports by Ross (2014) who identified no differences in reach distance between injured and uninjured participants on the SEBT, but did identify that injured participants exhibited greater hip flexion, trunk rotation and pelvis rotation to achieve maximum reach excursions. Ross (2014) suggested that these movement dysfunctions of the hip and trunk observed during the SEBT could be clinically relevant for the rehabilitation and prevention of chronic ankle instability. Although the composite scores reported for the SEBT quality assessment by Ness (2015) presented ‘fair’ to ‘moderate’ inter-rater agreement (W=0.64-0.73) there was no relationship between these and anterior reach distances reported between injured and uninjured participants. The authors hypothesized that
subjects who displayed fewer movement faults while still being at risk per SEBT scoring criteria may be unable or unwilling to move toward the limits of their stability, therefore decreasing the use of aberrant movement patterns.

A further example of a clinically meaningful movement screen that originated outside of the literature is the Landing Error Scoring System (LESS) (Padua et al., 2009). The LESS was designed to fulfil a specific clinical need for identifying knee injury (particularly anterior cruciate ligament) risk factors through a dynamic landing based task (Beutler, la Motte, Marshall, Padua, & Boden, 2009; M. C. Boling et al., 2009; Padua et al., 2009). The LESS is scored based on a count of 17 defined landing technique errors which are visually rated using video assistance. Since its initial presentation, reliability investigations have found the test to have ‘good’ to ‘excellent’ inter and intra rater reliability (Boling et al., 2005; Onate, Cortes, Welch, & Van Lunen, 2010; Padua et al., 2009). The LESS has also been adapted to be scored in real time and this too has been reported as a reliable test when scored by expert raters (Padua et al., 2011).

Most recently, the Triple Joint Flexion Test (TJFT) has been described by (Taingahue, 2015). The TJFT originated from clinical practice and was designed as a practical tool for identifying lower extremity movement errors in a range of functional lower extremity tasks. The TJFT has undergone considerable practical testing in developmental soccer and volleyball athletes prior to a recent formal investigation by Taingahue, (2015). To date, no rater reliability data has been reported for the TJFT.

2.2.7 Why does reliability matter?

As reliability is a pre-requisite requirement for both research and clinical application (Davidson & Keating, 2014), evidence for the reliability visual assessment of movement quality is essential. In the context of clients or individual athletes being managed by a single therapist, acceptable intra-rater agreement is important. In contrast, when screening groups of athletes, such as happens in pre-season screening of training squads, acceptable inter-rater agreement is arguably more important, as more than one rater may be involved in administering parts of a movement screening test battery, and raters need to be able to agree on findings so that further management is consistent.
2.3 What sources of variability within a rating are there?
All observation based movement screens have two inherent sources of rating variability, the athletes’ performance, and the raters’ subsequent analysis. Most published reliability studies have focussed on the variability observed within and between raters analysis. The variability of the athletes’ performance of the test is often excluded from analysis by using video, or included in the analysis without the required information to interpret the findings with reference to the expected athlete variability of the test. In order to fully understand the overall reliability and clinical utility of an observation based movement screen, researchers must first understand both these sources of variability.

2.3.1 Sources of variability attributable to athlete performance
The variation in an athlete’s performance can be quantified using 3-dimensional (3-D) and 2-dimensional (2-D) analysis. Although 3-D is considered the ‘gold standard’, few authors have undertaken 3-D investigations of intra athlete movement variation and this is likely due to resource requirements around cost and technical expertise required. In general, research findings from 3-D (Ford et al, 2007; Whatman et al., 2011; Milner et al., 2011) and 2-D analysis (Munro et al., 2012; Levinger et al., 2007) of common lower limb movement screening tests suggest that there is little intra-athlete, within- and between session variability in movement tests. For example, using 2-D analysis ‘excellent’ within and between session reliability has been reported by authors investigating single leg landings (Munro et al., 2012) and for single leg squat tasks (Levinger et al., 2007). Using 3-D analysis, Milner, Westlake, and Tate, (2011) also reported ‘good’ to ‘excellent’ within-session and between-session variability during a stop jump landing. It is, however, important to consider that the variation in levels of reliability reported between studies could be attributed to differences in study design, such as the movement tests, measurement variables, and participants recruited into the study.

Athletes displaying highly skilled performances and a decreased risk of injury have been reported to also demonstrate higher performance variability (Harbourne & Stergiou, 2009; Seifert et al., 2014; Stergiou, Harbourne, & Cavanaugh, 2006; Wagner, Pfusterschmied, Klous, Duvillard, & Müller, 2012). However, this performance variability is often not accounted for in movement screening protocols. For example, a single LESS-RT score is based on eight items which are scored separately over four repetitions of the test movement (Padua et al., 2011). Similarly, the SEBT (Kinzey & Armstrong, 1998) require ratings to be
based on the ‘best repetition’ of each movement, despite several repetitions been performed which may show a range of performance levels.

2.3.2 Sources of variability attributable to the rater

Within any given movement screening protocol there are several variables which may influence rater variability and subsequently rater reliability. These include: the scoring system (made up of the scoring scale and rating criteria), the raters understanding and skill in administering the test, and the movements being evaluated in the test. Notably, the isolated effect each of these variables has on rater agreement has received little attention in the movement screening literature. Nonetheless, the details of each variable are reviewed here in an attempt to describe their functions in a movement screening protocol.

2.3.2.1 Scoring system

Within a given movement screen, the scoring system informs the rater on how to interpret the scored movements. The scoring system consist of a scoring scale, and a scoring method. A scoring scale is used to grade the quality of movement observed during the movement screen, and the scoring criteria defines what body segments the criteria should be applied too. Subsequently, the ability of raters to interpret and use the scoring criteria with minimal variation between ratings is an important component of rater reliability.

2.3.2.2 Scoring scale

Scoring scales are used to grade the quality of movement observed during a movement screen. Two types of scoring scales are described in the literature; ordinal, and dichotomous. In general terms, ordinal scales used in the movement screening literature consist of 3 or more categories which are used to score the quality of movement, for example, as ‘good’, ‘fair’, or ‘poor’ for a given criterion. This method of scoring tends to result in clinically sensitive and descriptive information about the movement dysfunction observed (Chmielewski et al., 2007). In contrast, dichotomous scales might be used to indicate the presence or absence of a movement dysfunction, such as “did the patella move medial to the big toe? Yes or no?” This dichotomous response is less detailed than the ordinal scales, but tends to produce higher levels of rater agreement as there are fewer categories from which raters need to agree and the ‘presence or absence’ task is inherently less prone to error than when multiple categories are available to the rater.
2.3.2.3 Rating Criteria

In a movement screening context, the rating criteria describe which body segments are analysed during a movement test. Four variations of rating criteria are described in the literature, these are: overall scores, segmental scores, multi segmental scores and composite scores. Overall scores are based on the rater's general impression about the quality of a performed movement and are often scored on an ordinal scale as ‘good’, ‘fair’, or ‘poor’ (Chmielewski et al., 2007; Poulsen & James, 2011; Weir et al., 2010). Segmental scores, and multi segmental scores, are based on the specific movement quality of one or more joints and are scored on either dichotomous (Ageberg et al., 2010; Ekegren et al., 2009) or ordinal scales (Porat, Holmström, & Roos, 2008; Tofte, Tillman, & Chmielewski, 2011). Lastly, composite scores are calculated by adding together multiple single segment scores. For example, on a dichotomous scoring scale a movement error might be scored a ‘1’ if present, and a ‘0’ if absent. If four segments are then scored for a movement (for example, a squat) and 3 errors are scored by the rater, a composite score of 3 out of a possible score of 4 is reported. In this way, an overall impression of movement quality is determined by the sum of the individual segments being rated (Ness et al., 2015; Park et al., 2013; Piva et al., 2006).

Although direct comparisons between rating criteria is difficult due to individual variations in study designs and methodologies, it is generally accepted that overall scoring and multisegmental scoring result in lower levels of rater agreement (Chmielewski et al., 2007; Whatman et al., 2012). The low levels of rater agreement are likely due to the inherent difficulty in analysing multiple segments during a movement that might be undertaken in a few seconds. In contrast to scoring of multiple segments, single segment scoring methods produce less clinically useful information, but higher levels of rater agreement (Ageberg et al., 2010; Ekegren et al., 2009) and this is probably due to the relative simplicity of rating a single segment. Acceptable levels of rater agreement are reported by authors using the composite rating criteria (Ness et al., 2015; Park et al., 2013; Piva et al., 2006), despite multiple segments being used to calculate the composite score. The information obtained from these rating criteria is also considered useful, as the composite score can be interpreted in reference to the overall quality of the movement (Ness et al., 2015; Park et al., 2013; Piva et al., 2006). However, few studies exist using this method and there are significant variations in study design and methodology between the existing studies.
2.3.2.4 Rater Training

Within a movement screening context, rater training is used to reduce the likelihood of raters misinterpreting the scoring criteria and subsequently introducing a source of variance in scoring (Lucas et al. 2010). Rater training in many rater reliability studies of movement screening consists of at least 2-hours where raters are instructed on the specific scoring protocols for a movement screening tool. There are several authors who have used rater training when scoring functional lower extremity movement tasks (Park et al. 2013; Ness et al. 2015; Piva et al. 2006; Ekegren et al. 2009; Ageberg et al. 2010). Where rater training is not used, researchers cannot be confident that the level of agreement reported has not been influenced by confounding variables, such as raters misinterpreting the test and applying the scoring tool irregularly (Lucas et al., 2010). However, one prominent author in lower extremity screening literature argues that untrained raters are a ‘truer representation’ of the tests clinical application, as specific and detailed training may not occur in clinical practice (Whatman et al. 2012; Whatman et al. 2013; Whatman et al. 2015). Whatman has advanced the argument that a rater’s experience is a more relevant variable when reporting rater agreement. Raters of higher experience in Whatman et al., 2012 and 2013 studies produced higher levels of intra-rater agreement (experienced raters range AC1 = 0.74 – 0.88, novice raters range AC1 = 0.66 – 0.81) but similar levels of inter rater agreement (experienced raters range AC1 = 0.22 – 0.68, novice raters range AC1 = 0.28 – 0.71) when compared to rates of lesser clinical experience.

Researchers reporting rater reliability of movement screens should be careful when referring to ‘rater experience’, as experience may relate to one of two things: firstly, this term has been used to refer to rater’s clinical experience; and secondly, could refer to rater’s experience with the specific movement screen. It is logical that a rater who routinely uses a movement screen is more likely to score reliably as compared to a rater who has clinical experience, but is essentially a novice with the specific movement screen.

2.3.2.5 Movements included in the test

To ensure an athlete is prepared for the wide variety of movement demands placed on them during sport, the analysis of a range test movements should be used in a movement screening process, and the test movements should represent functional tasks involved in sport and physical activity (Cook et al. 2006; McKeown et al. 2014; Mottram and Comerford 2008). Within the range of functional tasks included in lower extremity movement screening, there are lower velocity movements (also known as ‘low threshold’ movements) such as squatting...
and lunging, and higher velocity (‘high threshold’) movements, such as jump landing (Padua et al., 2009) and hopping (Taingahue, 2015). Most often, lower velocity movements are used as these commonly appear in clinical practice, are easier to analyse in real time by visual observation and are similar to movements seen in sport and athletic training (Whatman et al. 2012; Piva et al. 2006; Park et al. 2013; Ness et al. 2015; Ageberg et al. 2010; Whatman et al. 2013; Chmielewski et al. 2007; Poulsen and James 2011). However, as higher velocity movements are characteristically representative of more demanding sport specific activities, such as landing and jumping, there is a strong rationale that higher velocity movement be included in movement screens. There is, however, a higher level of skill required to analyse higher threshold movements related to the speed and complexity of the movements (Padua et al., 2011; Whatman, Hume, & Hing, 2015). In an effort to mitigate the difficulty of rating higher velocity movements Padua et al., (2009) and Whatman et al., (2013) have reported using video assistance in the form of slow motion and/or repeat viewings to help raters score movements with greater reliability. However, the trade-off for accuracy is that video analysis is time consuming and clinicians often need to make decisions in real time. Ratings of higher velocity movements made without video assistance have been reported (Nilstad et al., 2014; Padua et al., 2011). However the testing procedures have either been over-simplified to a single segment (Nilstad et al., 2014) or are more complicated and have only been tested in small number of expert raters (Padua et al., 2011).

2.3.3 Summary of rater source of variability

In general, authors developing and investigating movement screens must make decisions about which variables are included in the movement screen knowing that there is an inevitable trade-off between reliability and clinical usefulness. In the interests of achieving clinically useful information with clinically acceptable levels of rater agreement it seems reasonable to suggest that raters score from a scale with as fewer categories as possible, but with as many segments to ‘rule in’ or ‘rule out’ the presence of dysfunction which requires further investigation. The scoring criteria should then consist of composite scores as this is likely to produce acceptable levels of rater agreement. To ensure a range of movement tasks are evaluated, both higher and lower velocity tasks should be included in the screen and ideally these should be analysed in as lesser resource dependant way as possible. Lastly, to ensure raters adhere to the scoring criteria with as minimal variation as possible so that a true measure of the test is achieved, sufficient rater training should be given so that variance
between and within raters can be attributed to sources other than insufficient skill or knowledge of the movement screening criteria.
2.4 A methodological critique of existing reliability studies.

2.4.1 Introduction

It is common for researchers to critically appraise the quality of studies which examine the reliability of clinical tools (Carlsson & Rasmussen-Barr, 2013; McCreesh, Crotty, & Lewis, 2015; Moloney, Hall, & Doody, 2012; Moran, Schneiders, Major, & Sullivan, 2015). A critical appraisal is useful for identifying the risk of bias a study has. It is important to identify the risk of bias present in a study so that appropriate generalisations and causal conclusions may be made. The object of this section is to therefore critically appraise the quality of key reliability studies specific to lower extremity movement screening.

2.4.2 Method of review

2.4.2.1 Appraisal tool

The Quality Appraisal of Reliability Studies (QAREL) checklist has been developed to critique the quality of a range of reliability studies (Lucas et al., 2013). Eleven items within the checklist are used to methodologically appraise specific components of reliability studies (Lucas et al., 2010). The eleven items cover: the spectrum of subjects and raters, rater blinding, stability of the trait, application and interpretation of the test, appropriate statistical analysis, and order effects of examination. The reliability and validity of the QAREL protocol has been established (Lucas et al., 2013; Lucas et al., 2010). The QAREL protocol has also been used to assess a wide variety of reliability studies including movement screens (Carlsson & Rasmussen-Barr, 2013; Moran et al., 2015), thermal quantitative sensory testing (Moloney et al., 2012), and acromiohumeral distance measurements (McCreesh et al., 2015). The operational definitions of QAREL have been adapted here to critique reliability studies specific to lower extremity movement screens and is similar to that described by previous authors investigating movement screening reliability (Moran et al., 2015).

2.4.2.2 Search strategy

Literature was identified using electronic database search of SPORT Discus, EBSCO Health, ScienceDirect and Google Scholar. Articles were found using combinations of keywords and subject terms including “Lower Extremity Movement Screen” and “rater reliability”. Hand searching of reference lists in retrieved articles was also used.
2.4.2.3 Eligibility criteria

The inclusion criteria were based on this review of the literature, which has identified factors most likely to result in high levels of rater agreement and clinical usefulness. To be included, studies must have consisted of at least 3 of the following 6 criteria: Multisegmental dichotomous scoring methods, a range of functional movement tasks, known test stability or expected variation in athletic performance, live or un-assisted real time video analysis, high velocity tests; and the design include rater training. Abstracts, conference abstracts, and dissertations were excluded because of uncertain peer review status.

2.4.3 Findings

Five studies meet the search criteria. These were the SEBT adaptation by Ness et al., (2015), The LESS-RT investigation by Padua et al., (2011), the movement screening battery by Whatman et al., (2012), and the forward and lateral step down investigations by Park et al., (2013) and Piva et al., (2006). Each study was reviewed using the QAREL criteria adapted for this specific purpose. The results of this appraisal are displayed in Table 1.
Table 1. Quality appraisal using the QAREL checklist

<table>
<thead>
<tr>
<th>Question</th>
<th>Whatman et al., 2012</th>
<th>Padua et al., 2011</th>
<th>Ness et al., 2015</th>
<th>Piva et al., 2006</th>
<th>Park et al., 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Was the test evaluated in a sample of subjects who were representative of those to whom the authors intended the results to be applied?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>2. Was the test performed by raters who were representative of those to whom the authors intended the results to be applied?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>3. Were raters blinded to the findings of other raters during the study?</td>
<td>U/C</td>
<td>U/C</td>
<td>U/C</td>
<td>Yes</td>
<td>U/C</td>
</tr>
<tr>
<td>4. Were raters blinded to their own prior findings of the test under evaluation?</td>
<td>U/C</td>
<td>U/C</td>
<td>N/A</td>
<td>Yes</td>
<td>U/C</td>
</tr>
<tr>
<td>5. Were raters blinded to the results of the accepted reference standard or disease status for the target disorder (or variable) being evaluated?</td>
<td>N/A</td>
<td>N/A</td>
<td>Yes</td>
<td>U/C</td>
<td>N/A</td>
</tr>
<tr>
<td>6. Were raters blinded to clinical information that was not intended to be provided as part of the testing procedure or study design?</td>
<td>U/C</td>
<td>U/C</td>
<td>Yes</td>
<td>U/C</td>
<td>U/C</td>
</tr>
<tr>
<td>7. Were raters blinded to additional cues that were not part of the test?</td>
<td>U/C</td>
<td>U/C</td>
<td>U/C</td>
<td>U/C</td>
<td>U/C</td>
</tr>
<tr>
<td>8. Was the order of examination varied?</td>
<td>Yes</td>
<td>N/A</td>
<td>N/A</td>
<td>Yes</td>
<td>N/A</td>
</tr>
<tr>
<td>9. Was the stability (or theoretical stability) of the variable being measured taken into account when determining the suitability of the time-interval between repeated measures?</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>10. Was the test applied correctly and interpreted appropriately?</td>
<td>Yes</td>
<td>Yes</td>
<td>U/C</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>11. Were appropriate statistical measures of agreement used?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Notes: U/C = unclear, N/A = not applicable
2.4.3.1 Considerations of external validity

External validity is the extent to which results can be generalized to other situations external to the study (Davidson & Keating, 2014; Lucas et al., 2010). The external validity of a movement screen directly corresponds with how applicable a test is in the ‘real world’. QAREL checklist items 1, 2 and 10 are used to critique external validity.

**QAREL Item 1: Was the test evaluated in a sample of subjects who were representative of those to whom the authors intended the results to be applied?**

The primary objective of reliability studies should be to evaluate the test in a sample who are representative of those in whom the test would typically be applied. This may refer to a specific group of atypical individuals, or a non-specific group which represents normal variation seen in clinical practice (Lucas et al., 2010). In either instance, subject characteristics should be defined so that readers can judge the appropriateness of making generalisations from the tested sample to the wider population. In relation to movement screening, the injured state, sporting background, and gender of the subjects or athletes sampled are important characteristics that should be described. All five studies included in this review met this criterion.

**QAREL Item 2: Was the test performed by raters who were representative of those to whom the authors intended the results to be applied**

It is important that the sample of raters used in any reliability study is well defined to allow for appropriate generalisation of results (Lucas, et al., 2010). Similarly, it is important that the sample of raters be representative of the population that the results are intended to be generalised to. Overall all five studies met this criteria, however it was noted that often only a small number of raters (typically n=1 to 2) with varying clinical experience and backgrounds were recruited into the studies. Limited sample sizes inherently reduce the overall generalizability of the results. Larger sample sizes, like the sample of 44 raters recruited by Whatman et al., (2012), are more likely to be an accurate reflection of the level of agreement expected in the sampled population. However, there are inherent logistical difficulties of larger sample sizes which can reduce the viability of conducting investigations. For example, Whatman et al., (2012) was un-able to provide an equal environment for all raters during repeat video analysis. The experienced raters in this study undertook repeat
measures in more favourable, un-guided conditions using personal computers, whereas less experienced raters were monitored and scored the tests communally.

**QAREL Item 10: Was the test applied correctly and interpreted appropriately**

Studies investigating the reliability of movement screens should make clear reference to the protocol used during data collection. Use of a standardised protocol improves the external validity of results as this is the protocol most likely to be adhered to in ‘real world’ use of the movement screen. Deviations or modifications from the standard protocol should be clearly stated by authors so that appropriate generalizations and interpretations of the results can be made. Rater training is often used to ensure raters equally interpret the rating protocol and subsequently score ratings with minimal irregularities (Ness et al., 2015; Park et al., 2013; Piva et al., 2006). The amount of training a rater has for a test is therefore an important consideration regarding the competency of a rater to interpret and implement a screen (Moran et al., 2015) and the generalizability of the results. Four of the studies met this criterion by providing sufficient information about correct testing procedures, interpretation and training, whereas Ness et al., (2015) indicated rater training was given but gave no description about training content or time spent. As Whatman et al., (2012) study design intended on measuring rater reliability in the presence of this potential bias, the testing procedure was considered to have been applied and interpreted appropriately.

### 2.4.3.2 Consideration for Internal Validity

Internal validity is the degree to which causal conclusions can be made from results. It requires the controlled reduction of bias through managing variables other than those intended to be investigated by authors. QAREL items 3 to 9 and 11 address internal validity with Items 3 to 7 addressing blinding.

**QAREL Items 3 to 7 (Blinding)**

Failure to blind raters to additional information beyond that being measured inflates reliability estimates and compromises the quality of a study. Five QAREL items directly critique the presence of blinding for raters own previous findings, the findings of other raters, subjects’ clinical information, and subject characteristics (Lucas, et al., 2010). As seen in Table 1 the authors of these reliability studies have not reported sufficient information regarding blinding to satisfy these criteria. The need to blind raters to all other findings is probably the most important, as inter and intra rater reliability measures are made on the basis that the rater is not influenced by their own or any other findings. Of the five studies, only
Piva et al., (2006) satisfied the criteria for blinding raters to all other findings, and although it is probable that the remaining authors blinded raters to all other ratings, they failed to provide sufficient evidence in reporting of blinding. Small reflective markers are potential additional cues which are often discussed in the literature (Ness et al., 2015; Whatman et al., 2012). Reflective markers are often placed on athletes over anatomical landmarks and are often used for additional research purposes, such as movement variability investigations (Taingahue, 2015; Whatman, Hing, & Hume, 2011). Although these markers are visible during movement analysis and scoring, it has been suggested that the small size of the markings is likely to cause only negligible effects on the reliability of ratings (Whatman et al., 2012) and that these landmarks would normally be palpated and identified by raters in clinic (Ness et al., 2015).

**QAREL Item 8: Was the order of examination varied?**

The order of examination should be varied to control against bias introduced in one of two ways: 1) order effects related to athlete performance and fatigue; or 2) raters being influenced by cognitive fatigue or cues left by previous raters (Moran et al, 2015). Details of randomisation should be made apparent to readers so that appropriate generalizations may be made. Authors of movement screens have several options for controlling these biases. Either all raters observe the movement at the same time in live or video playback (Ness et al., 2015; Padua et al., 2011; Park et al., 2013; Whatman et al., 2012), or one rater views a live movement and is in turn followed by the next rater (Piva et al. 2006). In the latter example, the rater order should be varied, so that a given does not undertake ratings in the same order each time.

**QAREL Item 9: Was the stability (or theoretical stability) of the variable being measured taken into account when determining the suitability of the time-interval between repeated measures?**

Reliability can only be attributed to agreement among raters if the stability of the test variable is known and controlled. If the stability of the test is not known, then the theoretical stability of the test should be considered when designing the study. Within movement screening, this means the within and between session variability of the test must be reported and controlled for when measuring and interpreting rater reliability. If the expected variability of an athlete’s performance of the test is not known, video recordings can be used so that raters view and score the same movement during any repeat measures. Only one series of studies in the movement screening literature has investigated variability of an athlete’s performance.
using 3-D motion analysis (Whatman, Hing, & Hume, 2011), followed by the reliability of rater’s analysis using video rating of the same lower extremity movement screen (Whatman et al., 2012). As discussed previously, there were some limitations in Whatman et al.’s (2012) study; however, the results of the previous investigation of movement variability mean that the findings can be interpreted with reference to the expected movement variability associated with the test. By contrast, the theoretical stability of Piva et al. (2006) step down task was not reported and shared live or real time video analysis for ratings was not used. In this study one rater assessed a live step down task, and was followed by a second rater who re-assessed another repetition of the live movement. Although the order of examination was then varied the theoretical stability of the movement was not controlled. Variation in the athlete’s performance of the test may therefore have led the raters to score non-identical movements.

QAREL Item 11: Were appropriate statistical measures of agreement used?

There are well-accepted statistical conventions for analysing and reporting rater- reliability within the field of movement screening. Intraclass correlation coefficients (ICCs) should be utilised to analyse continuous data, such as composite scores. For estimates of reliability of ordinal data the kappa family including: Fleiss’ Kappa for overall agreement between multiple raters; Cohen’s Kappa for paired ratings such as inter-rater reliability; estimates of reliability should be used. Kappa like statistics, such as the first order coefficient (AC1) described by Gwet (Gwet, 2008), should be used when there is a high level of homogeneity of raters, or rater bias is thought likely. Kappa is sensitive to trait prevalence (eg the proportion of athletes with movement dysfunction for a given test) which can lead to the ‘paradox of kappa’ where there is a high percentage agreement but low kappa (Feinstein & Cicchetti, 1990). Gwet, (2008) suggested that the ability of kappa to reflect agreement diminishes considerably as prevalence gets closer to 0 or 100%. If trait prevalence is around 50% both the kappa and AC1 statistics perform alike. Gwet (2001) suggested the AC1 can be interpreted in a similar manner to the kappa coefficient based on previous research by Landis and Koch, (1977): 0.01 – 0.20 = slight; 0.21 – 0.40 = ‘fair’; 0.41 – 0.60 = ‘moderate’; 0.61 – 0.80 = good; 0.81 – 1.0 = ‘almost perfect’.

2.4.4 Summary

This critical appraisal of lower limb movement screening reliability studies identified several risks for bias in five key research studies. Namely, incomplete descriptions of rater blinding
was identified as a common weakness amongst the studies. It is important for authors to describe rater blinding so that researchers and clinicians may understand the degree to which causal conclusions can be drawn from results. By contrast, most measures of external validity were adequately described by the authors.
2.5 Overall summary of the existing literature to describe the rationale for movement screening and identify the area of research need

Regular physical activity contributes to the primary and secondary prevention of several chronic diseases, including cardiovascular disease, diabetes mellitus, cancer (colon and breast), obesity, hypertension, bone and joint diseases (osteoporosis and osteoarthritis), and depression (Blair & Brodney, 1999; Bouchard et al., 1994; Gibbs et al., 2014; Mishra et al., 2012; Puett & Griffin, 1994; Taylor et al., 2004; Wagener & Klein, 1999; Warburton et al., 2001). In New Zealand, an increase in government funding has led to agencies such as Sport New Zealand (formerly SPARC) to promote participation rates in New Zealand (Sport New Zealand, 2015). However, as a consequence of increasing physically active lifestyles an increase in the number of sport and physical activity related injuries is inevitable (Verhagen & van Mechelen, 2010) and data from New Zealand’s Accident Compensation Corporation reflects this pattern (Accident Compensation Corporation, 2015). Of injuries which occur from sport or physical exercise, many involve the lower extremity (Fong et al., 2012; Hussain, 2010; LaBella, 2007), and often lead to diminished short and long term detriments to health (Verhagen & van Mechelen, 2010).

There are two major types of injury associated with sport and physical activity. Although it is generally accepted that acute impact type injuries are unavoidable, overuse type injuries may be avoidable where potentially modifiable injury risk factors are controlled (Verhagen & van Mechelen, 2010). Neuromuscular control is a modifiable risk factor which has achieved significant attention in recent research (Brent, 2011; Comerford & Mottram, 2001; Lee & Powers, 2014; Mauntel et al., 2013; Mottram & Comerford, 2008; Noyes et al., 2005; Zazulak et al., 2007). There is a growing body of evidence linking movement dysfunction with increased risk of the lower extremity overuse injuries (Chuter & Janse de Jonge, 2012; Felson et al., 2013; Lee & Powers, 2014; McKeon & Hertel, 2008; Stearns et al., 2013).

‘Movement Screening’ has been proposed as part of physical examination for identifying movement dysfunction primarily through observational analysis, often using several different tests – sometimes referred to as a ‘test battery’ (Frohm et al., 2012).

Several lower extremity movement screens have been described in the literature, including those which have originated from clinical practice, such as the SEBT (Kinzey & Armstrong, 1998), the LESS (Padua et al., 2009) and recently the TJFT (Taingahue, 2015) and those which have originated in the literature, including Whatman et al., series of investigations (Whatman et al., 2011, Whatman et al., 2012 and Whatman et al., 2013) and investigations by Chmielewski et al., (2007) and Ekegren et al., (2009). However, there is a general lack of
A rigorous evaluation of lower limb movement screening literature, and as a consequence the limitations for clinical use are not well understood (Kivlan & Martin, 2012). In order for clinicians and researchers to better understand the limitations of a movement screen, the validity and reliability of the tests needs to be defined (Davidson & Keating, 2014). It is also important that a movement screen will likely produce clinically useful and reliable information about the athlete being tested and authors developing and investigating movement screens must make decisions about which variables are included in the movement screen knowing that there is an inevitable trade-off between reliability and clinical usefulness. Of the studies that have been identified as likely to produce clinically useful and reliable information (Ness et al., 2015; Padua et al., 2011; Park et al., 2013; Piva et al., 2006; Whatman et al., 2012), a risk of bias from insufficient rater blinding has been identified. The TJFT is a relatively new movement screening tool which originated as part of injury reduction strategies for adolescent Norwegian and Qatar football (soccer) players. It has been suggested that the two sources of rating variability within the TJFT could be explored separately in a two part study (Taingahue, 2015), and this could be similar to previous lower limb movement screening investigations by Whatman et al., (2011) and Whatman et al., (2012).

As Taingahue, (2015) has conducted an initial investigation of the TJFT by exploring the athlete variability within and between session scores an investigation of the inter- and intra rater reliability of the TJFT is warranted (See figure 2) and is presented in section II. In completing a rater reliability investigation of the TJFT, clinicians and researchers would have

![Figure 2: The research progression of the TJFT.](image-url)

(a) = Investigation by Taingahue, (2015)
(b) = Present study
a greater understanding of the TJFT’s clinical utility. Moreover, for any given clinical measure, the limits of validity are constrained by reliability (Zumbo, 2007), therefore; reliability is a pre-requisite requirement for both research and clinical application (Davidson & Keating, 2014).
References


Hoch, M. C., Staton, G. S., McKeon, J. M. M., Mattacola, C. G., & McKeon, P. O. (2012). Dorsiflexion and dynamic postural control deficits are present in those with chronic ankle instability. *Journal of Science and Medicine in Sport, 15*(6), 574-579.


Section 2: Manuscript

Note to reader:

This manuscript has been prepared in accordance with the instructions for authors for *Physical Therapy in Sport*. References in square brackets to Thesis Appendices are included for examination purposes and will be removed prior to submission for publication.
Inter-rater and intra-rater reliability of the Triple Joint Flexion Test (TJFT) scores in adolescent field and court athletes by novice raters

Darryl Keith Jenkins
Department of Community and Health Services
Unitec Institute of Technology
Private Bag 92025, Auckland 1142
1 Abstract

Background Despite the common use of lower limb movement screening tools, the limitations for clinical use are not well understood. There are two sources of variation associated with reliability measurements in movement screening. These are: (1) the athlete’s performance of the test; and (2) the observational ratings of the test movements made by the rater. The Triple Joint Flexion Test (TJFT) is a new movement screening tool based on work with development level adolescent court and field athletes. To date, one previous study has reported objective data of intra-athlete within- and between-session variability of the TJFT. To date, no objective data of observational ratings of the TJFT sub-test scores has been reported.

Design A repeated measures, test re-test design was used to investigate inter and intra-rater reliability using ratings made from video observation in real time.

Aims To establish the inter- and intra-rater reliability of TJFT scores in a sample of adolescent male court and field athletes.

Method Seven trained novice raters rated TJFT sub-tests on two occasions using real time video. The TJFT was administered in a sample of 17 adolescent court and field athletes and 182 ratings were made on during each occasion.

Results Inter-rater weighted agreement of the six TJFT sub tests by n=7 raters on day one and n=6 raters on day two demonstrated ‘moderate’ (AC2 > 0.41) to ‘substantial’ inter-rater agreement (AC2 > 0.61). Intra-rater weighted agreement of the six TJFT sub-tests by n=6 raters was mostly ‘moderate’ (AC2 > 0.41) and ranged from ‘fair’ (AC2 > 0.21) to ‘near perfect’ (AC2 > 0.81).

Conclusion The inter-rater and intra-rater agreement of TJFT sub-test scores have now been investigated, and found to sufficiently reliable. It is uncommon for both athlete and rater sources of variability to be identified for a lower limb movement screen. As such, clinicians may use the TJFT with a greater understanding of the expected variability from both the rater and the athlete. In the wider movement screening literature, reports of rater-reliability for trained novices scoring multisegmental movements in real time is unique. This is also the first scoring protocol to have reported acceptable levels of inter and intra-rater agreement for a higher threshold movement task under these scoring conditions

Keywords: Movement screening, Reliability, Injury, Lower extremity, Triple Joint Flexion Test.
2 Introduction

Neuromuscular control is a modifiable risk factor for overuse musculoskeletal injury which has achieved substantial attention in recent research (Brent, 2011; Comerford & Mottram, 2001; Lee & Powers, 2014; Mauntel et al., 2013; Mottram & Comerford, 2008; Noyes et al., 2005; Zazulak et al., 2007). There is a growing body of evidence linking movement dysfunction with increased risk of the lower extremity overuse injuries (Chuter & Janse de Jonge, 2012; Felson et al., 2013; Lee & Powers, 2014; McKeon & Hertel, 2008; Stearns et al., 2013), and ‘movement screening’ has been proposed as part of physical examination for identifying movement dysfunction primarily through observational analysis, often using several different tests – sometimes referred to as a ‘test battery’ (Frohm et al., 2012).

Several lower extremity movement screens have been described in the literature, including those which have originated from clinical practice, such as the Star Excursion Balance Test (SEBT) (Kinzey & Armstrong, 1998), and the Landing Error Scoring System (LESS) (Padua et al., 2009), and those which have originated in the literature, such as Whatman et al., (2011), Whatman et al., (2012) and Whatman et al., (2013). However, with the exception of Whatman et al.’s series of investigations there is a general lack of investigation of lower limb movement screening literature, and as a consequence the limitations for clinical use are not well understood (Kivlan & Martin, 2012). In order for clinicians and researchers to better understand the limitations of a movement screen, the validity and reliability of the tests needs to be defined (Davidson & Keating, 2014). Two sources of variation are associated with reliability measurements, firstly, the athlete’s performance of the test movements; and secondly, the ratings of the test movements made by the rater. The validity of a movement screen is often investigated once the variability of the test is established as the limits of validity are constrained by reliability (Zumbo, 2007). Therefore, establishing reliability is a pre-requisite requirement for both research and clinical application (Davidson & Keating, 2014).

The Triple Joint Flexion Test (TJFT) is a relatively new movement screening tool which originated as part of injury reduction strategies for adolescent Norwegian and Qatar football (soccer) players (Taingahue, 2015). The TJFT includes a battery of higher and lower threshold functional tasks intended to challenge athletes’ dynamic range of motion and stability. Ratings of the TJFT are made in real-time, are based on multiple body segments, and require minimal equipment which is advantageous for a range of clinical or team based
screening situations. Recently, Taingahue, (2015) explored the athlete variability within and between sessions for TJFT scores, however, no investigations of rater variability have been conducted. As it is essential that both sources of variability are investigated so that validity research may begin (Zumbo, 2007) an investigation of the inter- and intra rater reliability of TJFT scores is warranted. Therefore, the aim of this study was to investigate the inter- and intra rater reliability of TJFT scores.

3 Methods

3.1 Design
A repeated measures, test retest design was used to investigate inter and intra-rater reliability of TJFT scores made by 6 raters who observed video clips of TJFT sub-tests in two sessions separated by an interval of 14 days. A 14 day separation between tests is thought to be sufficient to minimise recall bias (Campbell, 1991) and this interval is common in the literature (Ekegren et al., 2009; Whatman et al., 2012). All participants gave written informed consent to participate in the study. The study was approved by the Unitec Research Ethics Committee (UREC 2013-1019).

3.2 Participants
There are two sets of participants in the study: the athletes, and the raters. The athletes performed the TJFT sub tests whilst the raters scored the athletes performances. The two sets of participants and the recruitment processes are described here.

3.2.1 Recruitment and description of athletes
As part of a prior investigation (Taingahue, 2015) a convenience sample of 17 adolescent male athletes (mean age 16.9 ± 0.9 years, mean height 182 ± 5 cm, mean body mass 77.4 ± 12.0 kg) were recruited from a secondary school court and field sports program. Inclusion criteria for the study were: (1) currently active in a sport season or having participated in a sport within the past calendar year; and (2) report maintaining a physically active lifestyle. Athletes were excluded if they: (1) had sustained an orthopaedic injury within the past six months, or (2) had a pre-existing cardiovascular, pulmonary, or metabolic condition which prevented them from participating in sport. In total ~900 videos were made of the athletes performing TJFT sub-tests. A sample of 182 videos was then purposefully selected to ensure that (1) a full range of TJFT scores, sub-tests and views was presented; and (2) video clips
were of acceptable quality for the viewing purposes of the study. Videos from all 17 athletes were included in the pool.

3.2.2 Recruitment and description of raters

A convenience sample of 7 raters (3 female, 4 male) undertaking a postgraduate degree in osteopathy (Auckland, NZ) were recruited. None of the raters had previous experience with the TJFT. All raters undertook 4 hours of training in two sessions over two days prior to data collection. Raters were recruited on the basis that they could commit to the appropriate time required to complete the study. All those who indicated a willingness to participate were invited to do so. Raters were blinded to all clinical information not intended to be part of the test, including athletic ability and injury status of the athletes. One rater dropped out before the second data collection day for scheduling reasons.

3.3 TJFT protocol

Full details of the TJFT protocol and rating criteria are reported elsewhere (Taingahue, 2015) [See Thesis Appendix 1]. In general terms, the TJFT consists of three movements and six movement sub-tests derived from frontal and sagittal views of each of the movements. The sub-tests are scored from a standardised set-up, scoring requirements and scoring criteria. Once the setup is achieved, the movement is analysed. If the movement meets the requirements, one (1) point is awarded. If the requirements are not met then zero (0) points are awarded and a rating of movement quality is not made. If the athlete does meet the requirements of a movement test they score one (1) point and their movement quality is rated according to a multisegmental dichotomous criteria (three frontal plane and two sagittal plane criteria for each sub-test). One (1) point is scored for each of the criterion satisfied resulting in a possible score ranging between one (1) to six (6) points for each sub-test, in which the requirements were met. For the purposes of the study, overall composite sub-test scores made from both sagittal and frontal view ratings were not used. Instead the scores for individual sub-test views (frontal or sagittal) were used to enable comparisons between sub-test views.

3.3.1 Visual assessment procedures

To ensure equal viewing angles during testing, raters were divided into two equal groups based on availability. Each group then attended two data collection days two weeks apart. For all tests, raters visually assessed movement quality based a modified TJFT protocol [See Thesis Appendix 2]. The modifications clarified issues which arose in rater training, such as
what constituted a change in foot position or heel lift. Notably, these clarifications did not change the overall visual assessment procedure. Videos were presented for assessment on a 50-inch high definition display in a dimly lit, quiet room. All raters simultaneously observed each video and no pausing, slow motion, or repetition was permitted. Raters worked independently without communicating with each other. Raters were given 30 seconds to rate each movement in real-time before the next movement was displayed. Raters were given 10-minute breaks every 45 minutes to minimise the influence of rater fatigue. The whole session was 120 minutes duration. Raters recorded their scores on the standardised TJFT data sheet. To reduce the risk of rater bias, the order of presentation was assigned randomly and raters were blinded to ratings made previously by themselves and other raters.

3.4 Data analysis
Raw data was extracted from the data collection sheets and tabulated in a spreadsheet and checked. The composite score for each sub-test view (frontal or sagittal) was calculated and used for subsequent data analysis. To estimate agreement within and between raters Gwet’s first order agreement coefficient (AC1) (Gwet, 2010) was calculated using Gwet’s first (AC1) and second order (AC2) agreement coefficients. The AC2 coefficient was calculated using linear weightings (set at 1, 0.75, 0.50, 0.25 and 0 in respect to the magnitude of disagreement) to account for slight differences in agreement which are not likely to result in different management plans and penalize larger discrepancies in agreement that would likely result in different treatment plans. Confidence intervals (95%CI) were constructed for all reliability estimates. Descriptors for magnitudes of agreement were adopted from Landis and Koch (1977): 0.01-0.20 = slight; 0.21-0.40 = fair; 0.41-0.60 = moderate; 0.61-0.80 = good; 0.81-1.0 = almost perfect. The minimum agreement coefficient to be considered clinically acceptable was operationally defined as AC1 or AC2 ≥ 0.4. This threshold is similar to that previously defined as acceptable for kappa coefficients (Moran et al., 2015). All analyses were undertaken using AgreeStat (2015.5, Advanced Analytics, Gaithersburg, MD, USA).
4 Results

4.1 Inter-rater reliability
Weighted agreement of the six TJFT sub tests scored by \( n=7 \) raters on day one, and \( n=6 \) raters on Day 2 demonstrated ‘moderate’ to ‘substantial’ inter rater agreement (See figure 1). The level of agreement was slightly higher on Day 2 than Day 1 and this may suggest the presence of a learning effect. Specifically, the level of agreement for hop and stick in the sagittal view (both days) and squat in the frontal and sagittal view (Day 2) achieved ‘substantial’ agreement amongst all raters. All other ratings achieved ‘moderate’ agreement across both days. Raters scores for all TJFT sub-tests where similar.

4.2 Intra-rater reliability between days
Intra-rater agreement was calculated for \( n=6 \) raters who attended both data collection sessions.

4.2.1 Frontal view
Weighted intra-rater reliability for frontal view sub-tests reached mostly a ‘moderate’ rating with a range of agreement from ‘fair’ to ‘near perfect’ (See figure 2). Specifically, hop and stick agreement ranged from ‘fair’ to ‘moderate’, single leg squat agreement ranged from ‘moderate’ to ‘substantial’ and squat agreement ranged from ‘fair’ to ‘near perfect’ (See figure 2).

4.2.2 Sagittal view
Weighted intra-rater agreement for all sagittal view sub-tests ranged from ‘moderate’ to ‘near perfect’. There were no differences in the range of agreement reached between sagittal view sub-tests (See figure 3).

4.3 Influence of weightings
Both weighted (AC2) and unweighted analyses (AC1) were used to determine the influence slight differences between scores on the levels of intra rater agreement. Sub-tests scored in the sagittal view mostly reached at least ‘moderate’ agreement with and without linear weightings (See figure 2&3). Following weighted analysis seven sub-tests scored in the frontal view went from ‘slight’ or ‘fair’ agreement to ‘moderate’ agreement (See figure 3). Specifically, for the squat sub-test raters 1 and 2 went from ‘slight’ to ‘moderate’ agreement and rater 4 went from ‘fair’ to ‘good’ agreement. For the single leg squat sub-test, raters 2 and 5 went from ‘fair’ to ‘moderate’ agreement and rater 1 went from ‘slight’ to ‘moderate’ agreement and rater 3 went from ‘slight’ to ‘good’ agreement (See figure 3).
Figure 1: Inter rater reliability of TJFT sub tests. A= Day one. B= Day two. Open squares = Frontal views. Closed squares = Sagittal views. Shaded area = Area of acceptable agreement.
Figure 2: Intra-rater reliability of TJFT sub-tests scored in the frontal view. (a) = Hop and stick (b) = Squat (c) = Single leg Squat
Shaded area = Area of acceptable rater agreement
Figure 3: Intra-rater reliability of TJFT sub-tests scored in the sagittal view. (a) = Hop and stick (b) = Squat (c) = Single leg Squat. Shaded area = Area of acceptable rater agreement
5 Discussion

5.1 Introduction
The TJFT was developed to meet the practical needs of movement screening amongst development-level athletes and sport teams (Taingahue, 2015). All observation based movement screens have two inherent sources of rating variability, the athletes’ performance, and the raters’ subsequent analysis. It has been suggested that the two sources of rating variability within the TJFT could be explored separately in a two part study (Taingahue, 2015). Previous investigations of the TJFT have identified the expected athlete variability within and between sessions (Taingahue, 2015). The aim of this study was to investigate the inter-rater and intra-rater variability of the TJFT sub-test scores. Trained novice raters scored real time video footage on two occasions. Overall, scoring of the TJFT sub-tests reached clinically acceptable levels of reliability (AC2 > 0.41). There were no substantial differences in scoring agreement between the sub-tests, or between testing days, which suggests the testing protocol is consistent for all sub-tests. This investigation is one of only two investigations into the TJFT. Moreover, in the wider movement screening literature, investigating reliability of scores attained by trained novice raters scoring multisegmental movements in real time, is unique and is not directly comparable to any previously reported studies. Indirect comparisons to similar movement screening studies are made here to give context to the present investigation.

5.2 Scoring protocol
Although no other lower extremity scoring protocol closely resembles the TJFT, multi segmental dichotomous scoring criteria have been described in the literature (Ness et al., 2015; Padua et al., 2015; Park et al., 2013; Piva et al., 2006; Whatman et al., 2012, 2013). Of these, three authors have reported on scoring methods and sub-tests similar to the movements and scoring included in the TJFT (Ness et al., 2015; Park et al., 2013; Piva et al., 2006). Ness et al., (2015) reported the inter rater agreement of a movement quality based scoring criteria for the SEBT as ‘moderate’ (W=0.66 - 0.73) for composite scores of 100 real time video ratings by n = 3 trained raters. Using step down tasks and identical scoring methods Park et al., (2013) and Piva et al., (2006) also reported similar levels of inter rater agreement (κ=0.80 and κ=0.67) to Ness et al., (2015) for live ratings by trained raters (n = 2 and 2). No intra-rater agreement was reported by any these three authors, although it seems reasonable to speculate that the level of intra rater agreement for these tests could be similar to the intra-
rater reliability of the TJFT, due to similarities in raters, scoring methods, and some sub-test movements. Conversely, these similarities may also be suggestive of the inter rater agreement expected to be achieved by the TJFT during live analysis. However, not all test movements used in these studies were similar to the TJFT sub-tests, and only the TJFT included a higher threshold movement. Lastly, Ness et al., (2015), Park et al., (2013) and Piva et al., (2006) all reported lower recruitment numbers (n = 2 to 3) and this makes sustainable generalizations to the TJFT difficult.

The level of intra rater agreement achieved by this study is similar or higher than that reported previously (Onate et al., 2010; Padua et al., 2009; Whatman et al., 2012, 2013). However, no previous author has reported intra rater agreement for strictly real-time video ratings. As video assistance is unlikely to be used clinically, due to time and resource constraints, the intra rater reliability of live ratings is important but significantly under reported in the literature. For example, Whatman et al., (2013) and Whatman et al., (2012) reported real-time viewing speed, but allowed the same repetition to be replayed multiple times. Similarly, Padua et al., (2009) and Onate et al., (2010) used slow motion with repeat viewings to make ratings of the LESS.

5.3 High-velocity movement sub-tests
To ensure an athlete is prepared for the wide variety of movement demands exposed to them during sport, the analysis of a range of functional movements should be included in a movement screening process (Cook et al., 2006a; McKeown et al., 2014; Mottram & Comerford, 2008). Most often low threshold movements such as squatting, single leg squat and step down tasks are used (Ageberg et al., 2010; Chmielewski et al., 2007; Ness et al., 2015; Park et al., 2013; Piva et al., 2006; Poulsen & James, 2011; Whatman et al., 2012, 2013). However, higher threshold movements seem to be more representative of sport specific lower extremity functional tasks, such as landing and jumping, and should therefore be included in movement screens.

Authors of movement screens which have included higher threshold movements have reported a range of inter and intra rater scoring reliability (Ekegren et al., 2009; Onate et al., 2010; Padua et al., 2011; Whatman et al., 2013). However, this study appears to be the first where a high-velocity task (i.e. single leg hop and stick) is scored based on multiple segments during a single viewing using un-assisted real time video. Similar investigations by previous authors have used video assisted analysis (Boling et al., 2005; Padua et al., 2009, 2015;
Whatman et al., 2012, 2013), multiple views and repetitions to make a single rating (Padua et al., 2011), or single segment ratings (Ekegren et al., 2009; Whatman et al., 2013).

5.4 Rater Training
To ensure the TJFT was applied correctly and interpreted appropriately basic rater training was given to all raters prior to data collection. The training consisted of 4 hours of practical based instruction on TJFT scoring tool. Rater training is common in the literature (Hyong & Kim, 2014; Ness et al., 2015; Onate et al., 2010; Padua et al., 2009; Park et al., 2013; Piva et al., 2006). However, there are only a few authors who have used rater training when scoring multi-segmental triple joint flexion tasks (Ness et al., 2015; Park et al., 2013; Piva et al., 2006) and with the exception of the present study, none have reported intra-rater reliability measures. The level of inter rater agreement reported by these authors has been ‘substantial’ to ‘near perfect’ which is similar to that observed in the current study. ‘Near perfect’ inter rater agreement was reported by Park et al., (2013) by two raters after 5 hours of comprehensive training. The 2 hours of training reported by Piva et al., (2006) during a similar study resulted in lower ‘moderate’ levels of rater agreement.

5.5 Rating Procedure
All scores of the TJFT sub-tests were made during real time video observations. Although observations from video may not fully represent live observations (Swaine & Sullivan, 1999), the decision to employ video observation in this study was undertaken to control for athlete variability between rating days. Previous authors have controlled for theoretical stability by rating an initial observation in real time, and then using video based observation for repeat measures (Chmielewski et al., 2007; McKeown et al., 2014; Mischiati et al., 2015). This method of analysis is, however, subject to variation in rater environment and viewing angles which may influence rater scoring and reliability (Moran et al., 2015; Swaine & Sullivan, 1999).

5.6 Limitations
In design of study care was taken to ensure raters were blinded to any information not intended to be part of the test. Raters were also asked not to communicate with other raters or review previous ratings within and between sessions.

There are several limitations inherent in this study. Firstly, the results of the current study are generalizable only to raters with similar levels of experience to those studied here. Secondly, movement analysis using the TJFT was performed in adolescent male court and field athletes; therefore, generalising to other populations should be undertaken cautiously. In particular,
generalizing to female athletes should not be assumed, due to differences between genders in lower limb anatomy, predisposition to injury, and factors that predict injury risk (Zazulak et al., 2007; Gribble et al., 2012). Thirdly, small reflective markers were placed over the skin of anatomical landmarks on the athletes and where visible to the raters. It is possible that these markings had some influence on the raters’ ability to identify landmarks during the visual analysis procedure. Reference markings are not uncommon in the literature (Ness et al., 2015; Whatman et al., 2012, 2013) and previous authors have suggested that small reference markers are unlikely to influence overall rater agreement (Whatman et al., 2012).

5.7 Implications for future research
This is the second investigation of the triple joint flexion test and follows one of three recommendations proposed by Taingahue, (2015). In conjunction with the remaining two recommendations by Taingahue, (2015) future specific recommendations for continued investigation and development of the TJFT include:

1. Investigating the sensitivity and/or rater reliability of TJFT scores on athletes with and without specific musculoskeletal pain condition, such as patellofemoral pain syndrome. This investigation could be conducted under live or real time viewings as described by previous author’s (Park et al., 2013; Piva et al., 2006)

2. Investigating inter rater agreement of TJFT sub-test scores during live scoring conditions. The observation conditions could be similar to Park et al., (2013) where raters viewed the movements at the same time and from identical as possible viewing positions.

3. Investigating the inter- and intra-rater reliability of TJFT sub-test scores on female athletes. This investigation could be similar to the present study.

These recommendations are intended on further exploring the external validity of the TJFT. As such, the recommended studies should be carried out in as close to clinically identical method as possible. Ultimately, these studies are part of the rigorous testing process required to develop the TJFT as a validated PPE tool for mitigating injury risk and monitoring recovery. Moreover, for any given clinical measure, the limits of validity are constrained by reliability (Zumbo, 2007) therefore; reliability is a pre-requisite requirement for both research and clinical application and should therefore be investigated further (Davidson & Keating, 2014).
6 Conclusion

The TJFT is a lower extremity movement screening tool designed to meet the practical needs of screening development-level athletes and sport teams. A previous investigations have found that the intra-athlete, between-session reliability of the TJFT using 2-D video analysis, as adequate for clinical monitoring of lower limb function in development-level, adolescent athletes (Taingahue, 2015). The inter-rater and intra-rater agreement of TJFT sub-test scores have now been investigated, and found to sufficiently reliable. It is uncommon for both athlete and rater sources of variability to be identified for a lower limb movement screen. As such, clinicians may use the TJFT with a greater understanding of the expected variability from both the rater and the athlete. In the wider movement screening literature, reports of rater reliability for trained novices scoring multisegmental movements in real time is unique. This is also the first scoring protocol to have reported acceptable levels of inter and intra-rater agreement for a higher threshold movement task under these scoring conditions.
Conflict of interest

None declared.

Ethical approval

Ethical approval for the study was granted by the Unitec Research Ethics Committee (UREC Approval No.: 2013-1019)

Funding

None declared.
7 References


8 Section 3 Appendices

Appendix 1: Description of TJFT

Appendix 2: TJFT standardised score cards

Appendix 3: Ethics documents: Consent form, information sheet, approval letter
Appendix 1: Description of TJFT

The TJFT lower extremity movement screening protocol consists of three lower extremity functional movements which challenge both mobility and motor control using higher and lower velocity movements. The three movements are; Squat, single leg squat (SLS), and single leg hop and stick (HS). The test requires that 3 repetitions be observed in the frontal and sagittal view. An assessment of each repetition is made based on the athlete's ‘end position’. The end position is described by a three second isometric contraction that occurs when the athletes reaches their end of range for the movement. For example, during the squat movement the end position is considered when knee flexion ends and the bottom of the thigh musculature is parallel with the floor.

Each movement has a similar set of five scoring criteria and are based on the multi segmental alignment observed during the end position. Each repetition is scored on a dichotomized scoring scale where a score of 1 is given if the criterion is achieved, and a score of 0 is given if the criterion is missed. A composite score is calculated based on the scoring of each segment criteria. A total of three points are available for each movement in the sagittal view and four in the frontal view. If the athlete fails to reach the specified end position or hold it for 3 seconds, or there is a change in their base of support (foot alignment/ heel raise), an automatic score of zero for that repetition is given.
### Appendix 2: TJFT standardised score cards

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Score</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>DL</td>
<td></td>
<td>3) Isometric hold in bottom position</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2) Bottom of high horizontal to floor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1) No heel lift or change in foot position</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Requirements:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3) Big toe &gt; 5cm lateral of heel</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2) Patella lateral to big toe RIGHT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1) Patella lateral to big toe LEFT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Relocation to foot</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pelvis level and central in relation to feet</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pelvis level and maintained lumbar lordosis positioned</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No = 1 / Yes = 0</td>
</tr>
<tr>
<td>Notes</td>
<td></td>
<td>Yes = Rate criteria</td>
</tr>
<tr>
<td>All</td>
<td></td>
<td>No - Do not rate</td>
</tr>
</tbody>
</table>

### Double Leg Squat (DL)

### Triple Joint Flexion Test (TJFT)
<table>
<thead>
<tr>
<th>Criterial</th>
<th>Left</th>
<th>Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pelvis Level</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Pelvis Level</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Paralle to Big Toe</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Paralle to Big Toe</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Trunk Maintain Vertical</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Trunk Maintain Vertical</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Pelvis Level</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Pelvis Level</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>No lateral sway</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>No lateral sway</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Frontal View</td>
<td>Score</td>
<td>Left</td>
</tr>
<tr>
<td>Frontal View</td>
<td>Score</td>
<td>Right</td>
</tr>
<tr>
<td>No = No Rate</td>
<td>Yes = Rate Criteria</td>
<td>Yes = Rate Criteria</td>
</tr>
<tr>
<td>Left</td>
<td>Right</td>
<td>Score</td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Lumbar lordosis maintained</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>No</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Pelvis level</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>No</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Parallel lateral to big toe</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>No</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>All requirements met</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td>No = 0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes**
- Do not rate
- Rare criteria

**Set-up:**
- 1) Test foot planted straight ahead
- 2) Free leg in front of body and flexed at hip and knee
- 3) Arms horizontal in front of body
- 4) No touch downs / brace on free foot

**Requirements:**
- 2) 90 degrees of knee flexion through 1 cm
- 3) Isometric hold in bottom position
- 4) No heel lift or change in foot position

**Single Leg Squat (SLS)**
**Tripod Joint Flexion Test (TFT)**
Participant consent form (Raters)

The reliability of a four-test movement screening protocol for visually assessing lower limb function in athletes

Name of Participant: ____________________________________________________________

I have seen and read the information sheet for raters taking part in the project titled “The reliability and validity of a four-test movement screening protocol for visually assessing lower limb function in athletes” and have had the opportunity to discuss the project with Matiu Taingahue or Rob Moran.

I understand that I am volunteering to partake in this study of my own volition, and I may withdraw at any time up to the completion of the data collection aspect of the research project.

I understand that my participation in this project is confidential and that no material that could personally identify me will be used in any reports on this project.

I understand that I can see the finished research document.

I have had enough time to consider whether I want to take part and acknowledge that any data collected during the study will be stored securely so that only the researchers may access them.

Participant Signature: ___________________________ Date: ________________

The principal researcher for this project is Darryl Jenkins and principal supervisor is Rob Moran:

Darryl Jenkins  
Tel: 022 012 6821  
Email: darrylj.999@gmail.com

Robert Moran  
Tel: 021 073 9984 or 815 4321 x8197  
Email: rmoran@unitec.ac.nz

The participant should retain a copy of this consent form

UREC REGISTRATION NUMBER: 2013-1019
This study has been approved by the UNITEC Research Ethics Committee from 20 December 2013 to 20 December 2015. If you have any complaints or reservations about the ethical conduct of this research, you may contact the Committee through the UREC Secretary (ph: 09 815-4321 ext 7248). Any issues you raise will be treated in confidence and investigated fully, and you will be informed of the outcome.
What should I bring?
During the teaching sessions, please come prepared to participate in practicing the movement screen. This means comfortable loose fitting or exercise appropriate clothing. As you will be rating movements and taking notes, a pen and highlighter is also required.

**Once I start can I withdraw from the study later?**

If you wish to withdraw from the study, you may do so for any reason up until the 24hrs after the conclusion of the final data collection stage. All personal information you provide will be treated as confidential and no material that could personally identify you will be used in any reports on this project.

**Who can I contact with any further questions?**

If you have any further questions about this research please feel free to contact one of us:

**Principal Researcher:**
Darryl Jenkins  
Tel: 022 012 6821  
Email: [Darrylj.999@gmail.com](mailto:Darrylj.999@gmail.com)

**Research Supervisor:**
Robert Moran  
Tel: 021 073 9984 or 815 4321 x8197  
Email: [rmoran@unitec.ac.nz](mailto:rmoran@unitec.ac.nz)

**UREC REGISTRATION NUMBER: 2013-1019**

This study has been approved by the UNITEC Research Ethics Committee from 20 December 2013 to 20 December 2015. If you have any complaints or reservations about the ethical conduct of this research,
you may contact the Committee through the UREC Secretary (ph: 09 815-4321 ext 7248). Any issues you raise will be treated in confidence and investigated fully, and you will be informed of the outcome.
Note to reader:
Ethics was granted to Matiu Taingahue as he was initially the only lead researcher for this study. I was added as a lead researcher on the same ethics contract at a later stage.
Unitec
Institute of Technology

Full name of author: Daryl Keith Jenkins

Full title of thesis/dissertation/research project: Inter-rater and intra-rater reliability of the triple joint flexion test (TJFT) scores in adolescent

Department of: Physiotherapy

Degree: MSc

Year of presentation: 2016

EITHER:

(1) I agree to my thesis/dissertation/research project being lodged in the Unitec Library (including being available for inter-library loan), provided that due acknowledgement of its use is made. I consent to copies being made in accordance with the Copyright Act 1994.

and

I agree that a digital copy may be kept by the Library and uploaded to the institutional repository and be viewable world wide.

OR:

(2) I wish to apply for my thesis/dissertation/research project to be embargoed for a limited period as per Academic Policy 12 Conduct of Student Research, Guideline 12/8.

Reason for embargo: ...........................................................................................................................................

Supervisor Approval: ...........................................................................................................................................

Dean, Research Approval: ...................................................................................................................................

Embargo Time Period: ...........................................................................................................................................

Signature of author: .................................................................

Date: .................................................................