Journal of Bodywork & Movement Therapies (2015) 19, 160—176

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/jbmt

PREVENTION & REHABILITATION: RELIABILITY RESEARCH

Interrater reliability of a Pilates movement-based classification system☆

Kwan Kenny Yu, PT, MPhty [a,b], Evelyn Tulloch, PT, MPhty [a], Paul Hendrick, PT, PhD, MCSP [c,*]

[a] School of Physiotherapy, University of Otago, New Zealand
[b] Queen's Rehabilitation Centre, Central, Hong Kong
[c] Division of Physiotherapy Education, University of Nottingham, Clinical Sciences Building, Nottingham, UK

Received 23 May 2014; received in revised form 19 August 2014; accepted 20 August 2014

KEYWORDS
Reliability;
Pilates;
Movement classification

Summary *Objective:* To determine the interrater reliability for identification of a specific movement pattern using a Pilates Classification system.
Method: Videos of 5 subjects performing specific movement tasks were sent to raters trained in the DMA-CP classification system.
Results: Ninety-six raters completed the survey. Interrater reliability for the detection of a directional bias was excellent ($_{Pi}$ = 0.92, and K_{free} = 0.89). Interrater reliability for classifying an individual into a specific subgroup was moderate ($_{Pi}$ = 0.64, K_{free} = 0.55) however raters who had completed levels 1—4 of the DMA-CP training and reported using the assessment daily demonstrated excellent reliability ($_{Pi}$ = 0.89 and K_{free} = 0.87).
Conclusion: The reliability of the classification system demonstrated almost perfect agreement in determining the existence of a specific movement pattern and classifying into a subgroup for experienced raters. There was a trend for greater reliability associated with increased levels of training and experience of the raters.

Introduction

Pilates is a discrete exercise approach which aims to target specific motor control patterns and re-educate mal-adaptive movement (Gladwell et al., 2006; La Touche et al., 2008; Rydeard et al., 2006). Traditional Pilates has been modified by physiotherapists for exercise interventions and

☆ This research was conducted at the School of Physiotherapy, University of Otago, New Zealand. At the time of the Kwan Kenny Yu was completing his Master of Physiotherapy Degree by dissertation under the supervision of Dr. Hendrick and Ms.Tulloch.

* Corresponding author. Tel.: +44 (0)115 8231827; fax: +44 (0) 115 846 8062.
E-mail address: paul.hendrick@nottingham.ac.uk (P. Hendrick).

http://dx.doi.org/10.1016/j.jbmt.2014.08.009
1360-8592/

List of abbreviations

CS	Classification System
DMA-CP	Dance Medical Australia Clinical Pilates
Email	Electronic Mail
HOP	Rebound Hop
K_{free}	Free-Marginal Kappa
LBP	Low Back Pain
MB	Movement-Based
$_{Pi}$	Percentage of Agreement
SHR	Single Heel Raise
4 PK	Four-Point Kneel

termed Pilates-based exercise (Lim et al., 2011; Rydeard et al., 2006; Tulloch et al., 2012). Dance Medicine Australia Clinical Pilates (DMA-CP), developed as a new Pilates-based approach, modified from traditional Pilates and provides specific rehabilitation exercises for patients with musculoskeletal disorders. The key feature of the DMA-CP concept is categorization of patients into specific subgroups by assessing individual's specific movement patterns (directional biases). The results from this assessment then inform decisions on management (Tulloch et al., 2012).

Subgrouping of patients who share similar clinical characteristics has been recommended to achieve improved patient outcomes (Long et al., 2004; McKenzie and May, 2003; Stolze et al., 2012). A number of movement-based classification systems have been proposed within the past two decades for subgrouping patients (Dettori et al., 1995; Karayannis et al., 2012). While some other movement-based classifications have undergone extensive research into their clinometric properties, (Dankaerts et al., 2006; Harris-Hayes and Van Dillen, 2009; Trudelle-Jackson et al., 2008) evaluation on the reliability of the DMA-CP system is limited despite its widespread use by physical therapists. A recent study of the DMA-CP found the reliability between two experienced raters was high (Tulloch et al., 2012). However, this research suggested that further study into the reliability and validity of the testing procedure was required within a larger sample size. Thus, results would have the potential to be more representative and therefore be generalized with a greater degree of confidence for clinical practice (Tulloch et al., 2012). Therefore the aims of this study were to investigate the levels of agreement within a large sample of raters working in clinical practice in using DMA-CP classification system to (1) detect the presence of a specific movement pattern(s) (directional bias); (2) categorize subjects into movement-based subgroups within a cohort of subjects with previous musculoskeletal injuries. The secondary aims were to investigate the effects of training and experience of raters on levels of agreement.

Method

Observational survey design: two groups of subjects participated in this study: video subjects and raters.

Video subjects

Five university staff and students (aged between 18 and 65 years) who had a present or history of back or lower limb musculoskeletal injury within 5 years prior to video recording were recruited.

Raters

The raters were recruited through the database of DMA-CP. All raters were required to hold a current physical therapy licensure and to have undertaken DMA-CP training within the previous 3 years. A total of 1320 eligible raters were invited into the study. Raters who agreed to participate but did not complete the survey were excluded from the study. All participants consented to participate in this study. This study was approved by the University of Otago Human Ethics Committee.

Procedure

The study involved 2 stages. The first stage involved developing a video recording of participants and survey preparation and the second stage involved administration of the online survey to the raters and data collection.

Video preparation

All videos were recorded by a musculoskeletal physical therapist (ET), who was highly experienced in DMA-CP directional bias assessment, and assisted by a postgraduate physical therapy student (KY). All recordings were undertaken in the same laboratory room using a standardized setting of mat position, camera and lighting set-up. The same digital video camera (Brand: JVC name of camera, Model: GZma30AA Hard Disk Camcorder) was used for each video recording.

Following informed consent, the video subjects were given standardized verbal instructions, to ensure the standardization of the movement tasks, including number of sets and order they were to be performed. KY demonstrated the following tasks before subjects performed them barefooted on a mat: (1) Single Heel Raise (SHR); (2) Rebound Hop (HOP); (3) Single Leg Kick (SLK); (4) Four Point Kneel (4 PK); (5) Roll ups. Figs. 1 to 5 illustrate the 5 tasks. These tasks followed the standardized protocol described by Tulloch et al. (2012) and are detailed in Appendix A. The standardized procedure for each movement task and recording are described in Appendix B.

Survey preparation

Completed videos were edited by the physiotherapist (KY) for the online survey. Each video was then edited to 5 min in length (KY). The 5 videos were then uploaded onto a password protected website which provided a media digital storage facility with a generated secure link for the video recordings. These links were then stored and embedded into an online survey detailed in Appendix C. Access to the online survey was provided by the principle investigator (KY) to participant raters. Raters were able to display a full screen image with video quality 640x480 pixels and 25 frames per second.

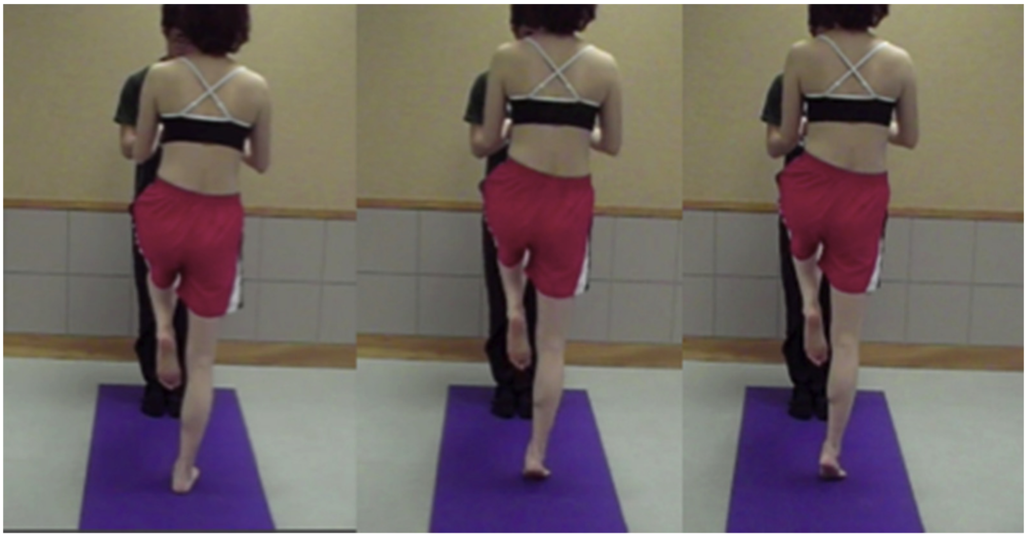


Figure 1 Sequence of Single Heel Raise from left to right.

Data collection procedure

Electronic mail (email) invitations were sent to all physiotherapists who had completed their most recent DMA-CP training within the previous three years; via the database of DMA-CP. The authors had no access to the database and as such no influence or bias on the selection of participant raters. Each potential rater received an information pack which included the study details. Confirmation of their willingness to participate was confirmed on receipt of their reply to the invitation (via email) to the PI during the 7-day rater recruitment period.

Each confirmed rater received a directional bias assessment manual (Appendix A) and instructions which included a web-link to the online survey. Raters were asked to view the video footage of the 5 subjects from the online survey in one viewing period if feasible. They were advised that they could view the footage as many times as necessary during the viewing period. Raters were asked to identify the presence of a directional bias for each subject (Yes/No), the subgroup to which each subject belonged and the key features that led them to their conclusion. Demographic data were also collected from the online survey. Raters were requested to submit the online survey within 14 days of receiving the survey. A follow up reminder was sent by email 7 days before the return deadline.

Statistical analysis

Frequencies of raters' demographic data were calculated and reported in percentages in Table 1. Free-marginal multirater kappa (K_{free}), as suggested by Randolph et al.

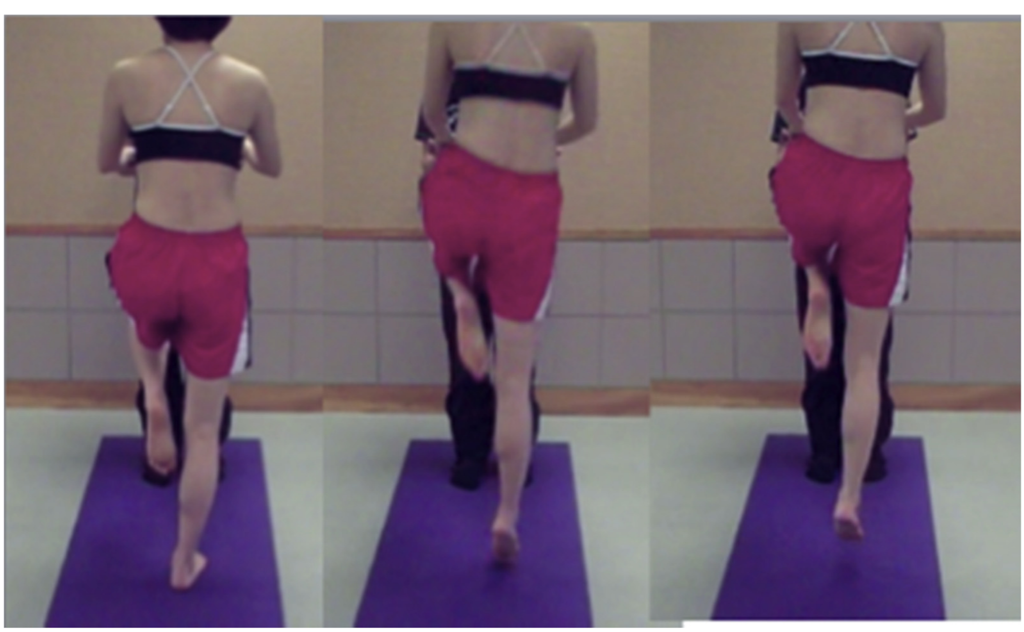


Figure 2 Sequence of Rebound Hop from left to right.

Figure 3 Sequence of Single Leg Kick from left to right.

Figure 4 Sequence of Four-Point Kneel from left to right.

(2005), was used to measure the overall interrater agreement for detecting the presence of a directional bias and for classifying each subject into a subgroup. The K_{free} is suggested for agreement studies in which the raters' distribution of cases into categories are not restricted (Randolph et al., 2005). It was therefore chosen for this study as marginal distributions were considered to be free and raters were able to assign cases to categories with no limits on how many cases could be placed into each category (Randolph et al., 2005). The K_{free} is also less likely to be influenced by bias or prevalence than fixed marginal kappa (Randolph et al., 2005).

The kappa calculation was undertaken in Microsoft Excel Version 14 (Microsoft Corporation) with the preset equation showed in Appendix D in an excel spreadsheet. An online Kappa calculation program (http://justus.randolph.name/kappa) was used to confirm the value of K_{free} (Randolph, 2008). The following guidelines were used for the interpretation of the K_{free}: less than 0 indicated poor agreement; 0.01 to 0.20 indicated slight agreement; 0.21 to 0.40 indicated fair agreement; 0.41 to 0.60 indicated moderate agreement 0.61 to 0.80 indicated substantial agreement; 0.81 to 1.00 indicated almost perfect agreement (Landis and Koch, 1977).

Results

One thousand three hundred and twenty electronic mail invitations were sent to practicing physical therapists through the database of DMA Clinical Pilates. Fig. 6 provides a flow diagram detailing rater recruitment and retention. One hundred and eight practicing physical therapists replied to the invitation and agreed to participate in this

Figure 5 Sequence of Roll ups from left to right.

PREVENTION & REHABILITATION: RELIABILITY RESEARCH

Table 1 Raters demographics.

Demographic variable		n, total = 96	(%)
Gender	Male	29	(30)
	Female	67	(70)
Age group (years)	18–29	29	(30)
	30–39	47	(49)
	40–49	12	(13)
	50–59	7	(7)
	60 or above	1	(1)
Practicing country	Australia	60	(63)
	Hong Kong, China	13	(14)
	Ireland	6	(6)
	New Zealand	3	(3)
	Singapore	4	(4)
	Sweden	2	(2)
	United Kingdom	4	(4)
	Others	4	(4)
Level of training	Level 1	8	(8)
	Level 1 & 2	33	(34)
	Level 1–3	10	(11)
	Level 1–4	45	(47)
Training of assessment	<6 months	28	(29)
	6 months–2 yrs	47	(49)
	2 year–5 yrs	19	(20)
	>5 yrs	2	(2)
Frequency in using the assessment	Everyday	48	(50)
	Once per week	26	(27)
	Once per month	7	(18)
	Once per year	5	(5)

Abbreviations: yrs, years.

study, representing an 8.1% return rate. Ninety-six questionnaires were returned of which 62 were fully completed.

Summary demographic data for the raters are presented in Table 1 and Table 2 respectively. The majority of the participants were female (70%), aged from 30 to 39 (49%), were practicing in Australia (63%); had completed all levels of DMA training (47%). Half the participants reported that they used the assessment every day.

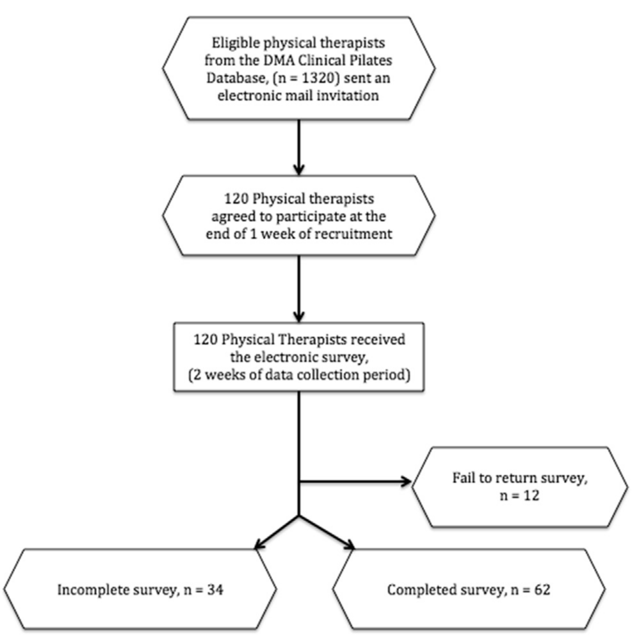


Figure 6 Flow chart of data collection procedure.

Table 2 Subjects Demographics, $n = 5$.

Variable	Values
Gender, male (%)	3 (60)
Age	33.85 ± 13.01
Report of previous LBP episodes[a] (%)	4 (80)
Report of previous musculoskeletal injuries in lower limb[b] (%)	2 (40)

Note: values are mean ± standard deviation (SD) unless otherwise indicated.
[a] Episode is a period of LBP that has limited the performance of daily activities for 24 h or more.
[b] Episode is a period of musculoskeletal injuries in lower limb that has limited the performance of daily activities for 24 h or more.

The overall agreement for raters who had completed data for all videos ($n = 62$) was $_{\text{Pi}} = 0.92$ for the detection of a directional bias and demonstrated 'almost perfect agreement' ($K_{free} = 0.89$) (Landis and Koch, 1977) (Table 3). There was a trend for increasing levels of agreement between raters with increased levels of training and experience, rising to $_{\text{Pi}} = 1.00$, $K_{free} = 1.00$ for raters who had completed levels 1–4 training and who also reported using the assessment daily (Table 4).

The percentage of overall agreement of raters ($n = 62$) to classify subjects into a directional bias subgroup was $_{\text{Pi}} = 0.64$, with $K_{free} = 0.55$ indicating 'moderate' agreement (Landis and Koch, 1977) (Table 3). Raters who had completed levels 1–4 training and who also reported using the assessment daily ($n = 21$) demonstrated almost perfect levels of agreement ($_{\text{Pi}} = 0.89$), and almost perfect reliability ($K_{free} = 0.87$) (Table 4).

Discussion

This study investigated the interrater reliability of the DMA-CP classification system within a cohort of DMA trained

Table 3 Distribution of raters' choices of video subjects in DMA-CP classification system.

		VS 1	VS 2	VS 3	VS 4	VS 5
Question 1: presence of directional bias	Yes	62	60	56	60	59
	No	0	2	6	2	3
Total valid sample for Question 1 for each VS		62	62	62	62	62
Question 2: directional bias subgroup	RE	56	1	4	0	0
	LE	4	1	49	4	2
	RF	1	41	1	7	7
	LF	1	14	2	47	50
Total valid sample for question 2 for each VS		62	57	56	58	59

Abbreviations: LE, left extension; LF, left flexion; RE, right extension; RF, right flexion; VS, video subject.

Table 4 Reliability of DMA-CP classification system of raters and associated level of training and experience of raters.

	Raters' categories	Percentage of agreement	Kappa coefficient
Detect the presence of a directional bias	Overall, $n = 62$	0.92	0.89
	Complete all training, $n = 35$	0.94	0.92
	Complete all training & frequent user, $n = 21$	1.00	1.00
Classify into directional bias subgroups	Overall, $n = 62$	0.64	0.55
	Complete all training, $n = 35$	0.78	0.73
	Complete all training & frequent[a] user, $n = 21$	0.89	0.97

*All training refers as 4 level of DMA-CP training. Each level consists of 16 h training.
[a] Frequent user referred to as raters using DMA-CP assessment everyday in clinical practice.

physiotherapists. To the authors' knowledge, this study is the first to conduct a web-based survey with embedded video recording in physical therapy research. The results found almost perfect agreement in detecting the presence of a directional bias (Agreement 92%, Kappa = 0.89) and overall moderate agreement (Agreement 64%, Kappa = 0.55) in classifying subjects into movement-based subgroups. The current study also found an improving trend of reliability with increased training and experience. There was substantial agreement (Agreement 78%, Kappa = 0.73) for experienced raters (completed 4 levels of training) who classified video subjects based on observed movement bias. Almost perfect agreement was reached (Agreement 89%, Kappa = 0.87) for experienced and frequent use raters. These results are similar to the previous study which found substantial agreement (Kappa = 0.76) for using the same subgrouping classification within two experienced raters for real subjects (Tulloch et al., 2012). The levels of agreement were similar when compared to the reliability of raters with similar training background in the current study.

Reliability in comparison to other movement-based CS

Studies on interrater reliability of movement-based classification systems demonstrate a range of reliability findings from moderate to almost perfect (Kappa from 0.61 to 0.96) (Dankaerts et al., 2006; Harris-Hayes and Van Dillen, 2009; Trudelle-Jackson et al., 2008; Vibe Fersum et al., 2009). However, a variety of methods have been employed to test interrater reliability of movement-based classification systems. Trudelle-Jackson et al. (2008); Harris-Hayes and Van Dillen (2009) tested the interrater reliability of a movement impairment-based system between 2 raters within a cohort of 24 and 30 LBP patients respectively. They reported similar levels of agreement to our study (Agreement 75%, Kappa = 0.61; Agreement 83%, Kappa = 0.75). However, it is difficult to compare the results directly as there are acknowledged differences between live and video reliability assessment (Swaine and Sullivan, 1999). Self-report and history information were recorded in Harris—Hayes and Van Dillens' study (Harris-Hayes and Van Dillen, 2009), and thus reliability results were not purely based on observation of movement. In addition, neither the effects of experience nor training were investigated in the above studies. Another study investigated the reliability of the O' Sullivan movement-based classification approach (Vibe Fersum et al., 2009). Four physical therapists, including the system developer and three others educated in the classification system, classified 26 LBP patients into subgroups. Almost perfect agreement (Agreement 86%, Kappa = 0.82) was found in this study when comparing the agreement between paired raters. Dankaerts et al. (2006) investigated the reliability of a subjective and video assessment patient subgroup classification. There was almost perfect agreement (agreement 97% and $K = 0.96$) between two expert clinicians when classifying 35 LBP patients into various movement-based directional patterns. Twenty-five patients were then videotaped and classified into directional groups by 13 other therapists based on the video and subjective complaints of the patients and substantial agreement was found between raters (agreement 70% Kappa = 0.61). It is evident from these results that reliability within movement-based classification assessment is affected by a number of factors including: the number of raters, the training of the raters, the number and type of subjects, the additional information given and whether the testing is conducted live or by video assessment.

Effect of training/experience on reliability of assessment

The results of the current study showed a clear pattern of improved reliability in Table 4 associated with more advanced levels of training and rater experience (frequency) in using the assessment. Swaine and Sullivan (1999) proposed that the amount of clinical experience and formal instruction in the examination procedure and classification rules are necessary prerequisites to improve reliability. The result from our study supports the importance of practice and experience in enhancing the reliability of movement-based assessment. The effects of training and experience have been previously investigated (Dankaerts et al., 2006; Fritz et al., 2000; Vibe Fersum et al., 2009). Vibe Fersum et al. (2009) found familiarity to the classification system increased the reliability for classifying LBP patients into subgroups (<100 practicing hours: Kappa = 0.66, >100 practicing hours: Kappa = 0.90). Dankaerts et al. (2006) also demonstrated improved reliability with familiarity of training (moderate familiar clinician: Kappa = 0.55, very familiar clinician: Kappa = 0.71; expert clinician: Kappa = 0.96) (Dankaerts et al., 2006). However, in that study familiarity was not specifically defined. Conversely, Fritz et al. (2000) found little difference in the reliability of the McKenzie classification system between students and practicing therapists, and between therapists with more or less clinical experience. The current study used frequency of clinical use of the assessment tool as the proxy measure of experience or familiarity with the assessment, which we believe better reflects clinical practice. Furthermore, raters were classified based on their level of training, which was standardized in content and training hours (16 h for each level of training). Although each movement-based classification systems differs, the results suggest reasonable reliability and validity for these systems which are dependent upon clinical training and expertise of the raters.

The current study standardized the setting, rater and participant instructions, the order of movement and the set of movements being recorded in the video, in order to lower potential sources of error caused by within-subject variability (Eastlack et al., 1991; Smits-Engelsman et al., 2008; Swaine and Sullivan, 1999). Unlike live assessment, subjects maintain consistency of performance in the movement tasks for video assessment. Video recording provides an additional advantage with the ability to capture observable behavior for later analyses (Gross and Conrad, 1991). However, it is acknowledged that video assessment differs from clinical assessment in a number of ways. Clinical assessment involves gathering a large amount of information including verbal feedback (individual's subjective feeling in movement tasks), proprioception feedback and audio feedback from the patient. Moreover, clinical and video assessment involves different viewing angles for evaluation of a subject's movement (Xu et al., 2011). The effect of the above differences between video and live assessment on reliability results is not clear. Swaine and Sullivan (1999) compared the interrater reliability assessment between video and live assessments on motor function within a cohort of head injury patients. The study showed differences in the scoring for the same rater in performing live assessment and video assessment. The authors found that raters required specific training in the scoring of videotaped observational data to maximize reliability (Swaine and Sullivan, 1999). This result may help to explain the overall moderate agreement which is lower than previous studies performed on live assessments (Swaine and Sullivan, 1999). Further research into the differences between interrater reliability testing employing live clinical assessment versus video assessment is therefore warranted to better investigate the constructive validity of this tool.

It is uncertain whether the classification categories encompass all possible movement biases and as such Tulloch et al. (2012) suggested variation of movement patterns or other combinations of directional bias may be observed clinically. This could be reflected by some raters who suggested a directional bias in the video subject; however they did not choose a subgroup for the subject in the existing choices. In the current study raters were not given an option to write 'other' as a movement category which did not fit the assigned categories and this could be considered for further research in this field. It is likely that reliability would be adversely affected if other potential categories of movement bias existed and this requires validation in future studies.

Study limitations and future directions

A number of potential limitations need to acknowledged and discussed. Although the results demonstrated a full spectrum of movement biases, it was a small sample size of subjects compared with previous reliability studies (Dankaerts et al., 2006; Harris-Hayes and Van Dillen, 2009; Trudelle-Jackson et al., 2008; Vibe Fersum et al., 2009). A previous study suggested that a more representative sample size improves the diagnostic accuracy of results (Cook et al., 2007). However, the potential effect of a larger sample size on the reliability assessment for a classification system is not clear and requires further study. In the current study, 5 subjects were chosen for pragmatic reasons and it could be strongly argued that increased numbers would be a significant rater burden and preclude raters' participation.

The relatively low response rate (8.1%) introduces a potential selection bias. Such a response rate compares unfavorably to previous studies employing an online survey (25%–40%) (Cook et al., 2000; Dankaerts et al., 2006; Kaplowitz et al., 2004). Response rates for online surveys can be affected by a number of reasons including the length of the recruitment period, failure of the online survey reaching potential participants due to invalid email accounts as well as time commitment (Dankaerts et al., 2006; Evans and Mathur, 2005). However, there was a very high return rate (89%) for those participants who agreed to participate which was encouraging. Through email communications with raters, we found the main reason for drop out was technological variations, including internet connection and the configuration of the user's computer that led to difficulty in survey or video access.

It is acknowledged that the results of this study can only be related to the video assessment of directional bias in DMA-CP. It has been suggested that all clinical simulation including video recordings are unable to approximate the full dynamics of a real subject encounter (Toro et al., 2003).

The presence of the assessment manual helped to standardize the procedure for assessment conducted by raters during video viewing. The use of the manual may have influenced the observed reliability by potentially improving the agreement particularly amongst those raters with less experience in using the assessment system. However, we did not specifically ask whether raters used the manual to rate the directional bias and therefore are unable to confirm this hypothesis.

Testing on interrater reliability is an essential step for validation of a classification system for use in clinical practice (Dankaerts et al., 2006; Dettori et al., 1995). Future studies should investigate test-retest reliability of the DMA-CP assessment system and establish whether reliability of the assessment tool differs in patient populations and further and investigate whether movement pattern differences in patient groups differ compared to healthy controls. The research also offers the possibility to compare the reliability of patient evaluation undertaken by those trained in a shared movement practice or systems. Ultimately, research needs to investigate whether subgrouping and exercise interventions improve rehabilitation outcomes. Therefore, this study forms part of the validation process for the use of DMA Clinical Pilates assessment in clinical practice.

Conclusion

This study investigated the interrater reliability of DMA-CP directional bias assessment within a large cohort of trained physical therapists. Raters could reliably detect the presence of a directional bias in all subjects. Depending on the level of training and experience, moderate to almost perfect reliability was found in classifying a subject into movement-based subgroups. Levels of training and experience were key factors affecting the levels of agreement for this movement-based classification system. Further research is required to validate the assessment procedure within patient populations. The authors consider that the validation of this assessment tool is imperative in order to identify potential patient subgroups and to investigate the effectiveness of subsequent treatment.

Appendix A. Directional Bias Assessment Instruction Manual

Directional Bias Assessment Instruction Manual

This instructional manual presents details for the directional bias assessment which consists of 2 assessment exercises used for baseline assessment and reassessments, 2 extension bias exercises with the lumbar spine in extension and 1 flexion bias exercise (with spine in flexion). The exercises were adopted from the DMA Clinical Pilates course manuals. The instructional manual was modified from the protocol established by Tulloch (Randolph et al., 2005).

Directional Bias Assessment

Baseline assessment

Participants performed the following 2 exercises for baseline assessment measures to determine whether the participant's performance was superior on the left or right side.

1) Single heel raise (SHR)

Participants were instructed to:

- stand with feet together, then
- stand on 1 leg, place their hands on top of the assessors hands;
- rise up on to their toes as high as possible with knee straight;
- and return to the starting position;
- repeat this 5 times on each leg;

The assessors: monitored differences in the performance of SHR between sides by comparing:

1. differences in the weight/effort they could feel through the participant's hands;
2. differences in balance;
3. difference in height raised;
4. signs of fatigue during the 5 repetitions.

Appendix A Figure 1 Single heel raise.

2) Hop

Participants were instructed to:

- stand with feet together, then
- stand on 1 leg, place their hands on top of the assessors hands;
- hop as high and rhythmically as they could, treating the ground like a hot plate;
- repeat 5 hops on each leg;

The assessors: monitored differences in the performance of Hops between sides by comparing:

1. difference in the weight/effort they could feel through the participant's hands as they hopped;
2. rhythm of the hop;
3. height of the hop;
4. surface area covered while hopping.

Appendix A Figure 2 Hop.

Lateral directional bias assessment

Lateral bias was determined from the participant's performance during baseline assessment measures (the poorer performing side) and was compared to whether the participant had a left or right sided bias during the performance of 3 mat intervention exercises, 2 with the spine in lordosis (extension bias) and 1 with the spine in flexion (flexion bias).

1. lumbar spine extension and/or rotation;
2. effort required to lift the leg;
3. improvement in or worsening in control of the lumbar spine with repetition.

Appendix A Figure 3 Single leg kick (SLK).

Extension bias

1) Single leg kick

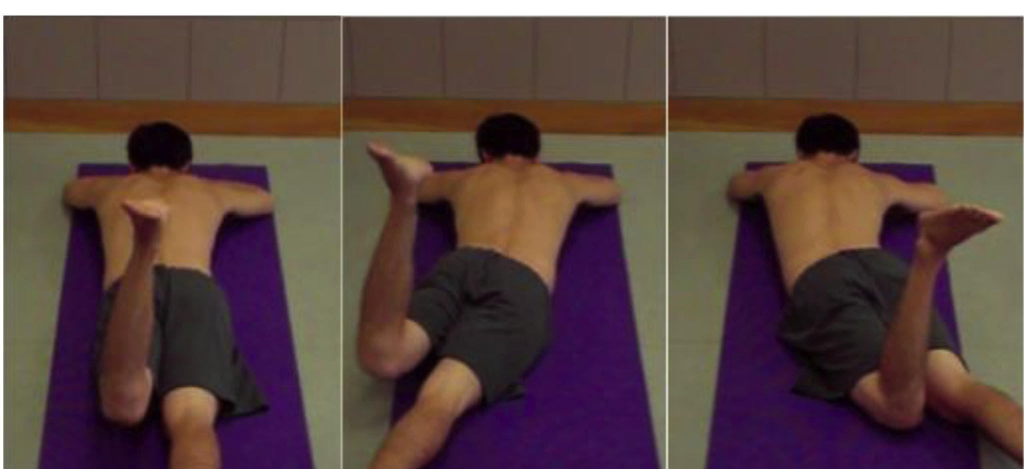
Participants were instructed to:
- lie prone with their feet together, legs in line with trunk place their hands under their head;
- bend their knee to 90°, draw their stomach up and;
- breathe in as they lifted their knee only 5 cm off the floor;
- breathe out as they straightened their knee and lowered their leg to the floor;
- repeat 10 times on each leg;
- rest for 1 min and repeat this sequence twice more.

The assessors: monitored differences in the performance of SLK between sides by comparing:

Retest anti-bias and Retest bias

Participants were then instructed to:
- take both legs to the side where it was observed the participant had performed the exercise with more stability (anti-bias);
- perform 1 set of repetitions in this position;
- take both legs across to the other side (bias);
- and perform another set of repetitions.

The assessors observed for any:
1. change in the participant's control and or effort to perform the exercise in the these 2 positions compared to the midline position.



Appendix A Figure 4 Directional bias assessment showing the left side with less stability, the mid, bias, and anti-bias positions during a single leg kick exercise.

2) Four point kneel

Participants were instructed to:

- get onto their hands and knees, place their hands together directly under their shoulders and their knees together directly under their hips, set scapulae in mid position and have their lumbar spine in a normal lordosis;
- keep their hips square as they breathe in and extend their leg out so it was parallel to the floor;
- breathe out as they brought their knee into their chest while maintaining a lordosis (disassociating hip flexion from lumbar flexion);
- return to the starting position;
- repeat 3 set of 10 repetitions on each leg with 1 min rest in between.

The assessors observed for:

1. the amount of lateral movement of the pelvis;
2. loss of lumbar extension and/or rotation control comparing 1 side with the other. Following 1 min rest this sequence was repeated twice more.

Flexion Bias

The participant was then tested for response to flexion bias. This was done by observing their performance during repeated roll ups. Participants were asked to repeat this exercise 5 times.

1) Roll ups

Participants were instructed to:

- lie supine, with feet together and with their arms overhead and their scapulae down;
- flex their shoulders to 90°;
- take a mid range breath in as they flexed their spine segmentally and rolled up into a long sitting position;
- segmentally roll down to supine;
- repeat 5 times.

After a short rest (up to 1 min) they were asked to repeat 5 more roll ups, unless their performance deteriorated. If they were still performing this exercise correctly the sequence was repeated.

Appendix A Figure 5 Four point kneel (4 PK).

Reassessment

From the SLK (increased movement observed) and 4 PK (increased difficulty on balance on supported leg) intervention exercises the assessors confirmed whether the left or right side was deficient. A reassessment of both the SHR and hop were performed. It was determined whether, compared to the baseline measure, the participant performed the SHR and hop with:

1. more or less stability;
2. more or less weight was felt through the assessor's hands;
3. rising higher or lower on to toes or hopped higher or less high;
4. fatigued any more or less quickly.

If the participant:

1. was unable to perform this exercise;
2. or found repeating 5 was too difficult;
3. or their ability to perform this exercise correctly and smoothly was deteriorating;

they were asked to stop and perform the roll down part of the exercise instead.

Participants were classed as flexion bias responders if:

1. they were able to perform either version of this exercise easily with good control;
2. could not perform 5 repetitions, but were able to improve with each set.

If the participants could not perform either version of the exercise or if their performance was deteriorating with repetition, they were deemed to be an extension bias responder.

Appendix A Figure 6 Roll ups.

To help confirm a left or right lateral bias both legs were taken to the better performing side (anti-bias), as determined from the above assessment and the participant was asked to perform a set of roll ups or roll downs. They were then asked to take their legs across to the other side and repeat the exercise for another set (bias). Any difference in performance was then noted between the 2 sides and the midline position.

Appendix A Figure 7 Roll ups with both legs taken to mid, bias and anti-bias position.

Reassessment

The SHR and hop tests were performed as above and any difference in the participant's performance was noted. If the participant was deemed to respond to a flexion bias and these reassessment tests were performed with:

1. better postural stability;
2. less fatigue towards the end of the hops or SHRs;
3. less weight felt through the assessor's hands;
4. the participant was able to hop higher or rise higher onto their toes;
5. a smaller surface area was covered while hopping

These parameters helped to confirm they were a flexion bias responder. If the reassessment tests (SHR and hop) were found to have deteriorated and the participant found the exercise of roll ups difficult or their form had deteriorated with repetition, this contributed towards confirming they responded to an extension bias.

Classification of directional bias assessment

Bias consists of 2 components:

- Left/right (lateral bias)
- Flexion/extension (sagittal bias)

The matched lateral bias was the side that was deemed to be deficient and requiring improved control. That is, there was more movement observed in the lumbar spine while the participant was performing the SLK exercise and lateral pelvis movement and or lumbar movement on supported leg during the 4 PK exercise. The flexion or extension matched bias was the direction that the participant had the better control, with a resulting better performance of the SHR and hop. Thus the participant-specific flexion or extension bias that was associated with improved performance was used to improve the lateral deficiency. The unmatched bias was opposite to the matched bias.

Appendix B. Video recording protocol

Part 1: Single heel raise (SHR)

- Instruction and demonstration to the participant of SHR on left/right.
- Participant's trial run (5 repetition).
- Rest for 10 s.
- Participant's performance of 3 sets of 5 repetitions on each side; video recording the last set on each side.

Part 2: Rebound hop (HOP)

- Instruction and demonstration to the participant of HOP on left/right.
- Participant's trial run (3–5 repetition).
- Rest for 10 s.
- Participant's performance of 3 sets of 5 repetitions on each side; video recording the last set on each side.

Part 3: Single leg kick (SLK)

- Instruction and demonstration to the participant of SLK of left/right with both legs in center.
- Participant's trial run (3–5 repetition).
- Rest for 10 s.
- Participant's performance of 3 sets of 10 repetitions on each side with 1 min rest in between; video recording the last 3 repetitions of last set on each side.
- Instruction and demonstration to the participant of SLK of left/right with both legs taking to the right side.
- Participant's performance of 1 set of 10 repetitions on each leg; video recording the last 2 repetitions of each leg.
- Instruction and demonstration to participant of SLK of left/right with both legs taken to the left side.
- Participant's performance of 1 set of 10 repetitions on each leg; video recording the last 2 repetitions of each leg.

Part 4: Four-point kneel (4 PK)

- Instruction and demonstration to the participant of 4 PK on left/right.
- Participant's trial run (3–5 repetition).
- Rest for 10 s.
- Participant's performance of 3 sets of 10 repetitions on each side with 1 min rest in between; video recording the last 3 repetition of last set on each side.

Part 5: Reassess SLK and HOP after extension exercise

- Participant's performance of 3 sets of 5 repetitions of SLK and HOP on each side, video recording the last set on each side.

Part 6: Roll ups

- Instruction and demonstration to the participant of roll ups with both legs in center.
- Participant's trial run (3 repetition).
- Rest for 10 s.
- Participant's performance of 3 sets of 5 repetitions on each side with 1 min rest in between; video recording the last set on each side.
- Instruction and demonstration to the participant of roll with both legs taken to the left.
- Participant's performance of 1 set of 10 repetitions with both legs taking to the left; video recording the last 2 repetitions.
- Instruction and demonstration to participant of roll with both legs taking to the right.
- Participant's performance of 1 set of 10 repetitions with both legs taking to the right; video recording the last 2 repetitions.

Part 7: Reassess SLK and HOP after flexion exercise

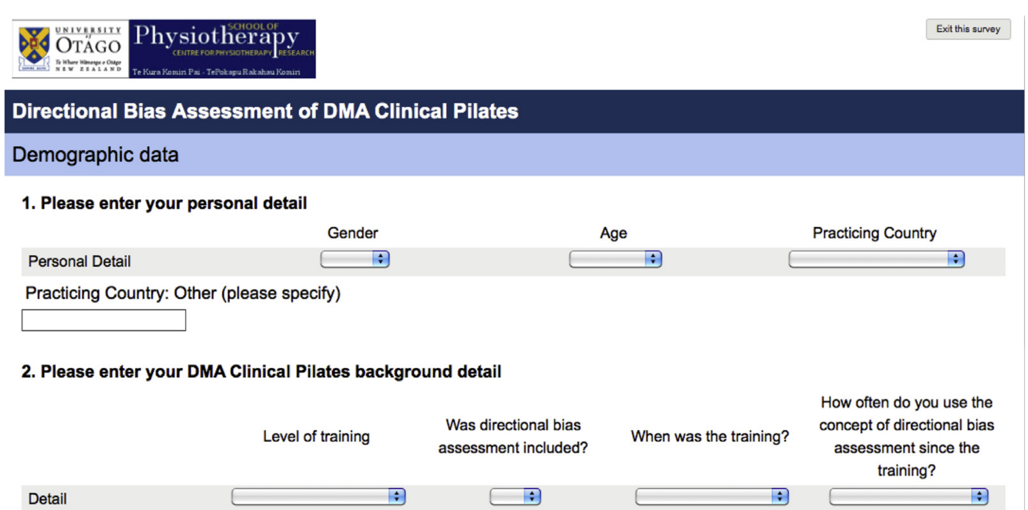
- Participant's performance of 3 sets of 5 repetitions of SLK and HOP on each side, video recording the last set on each side.

The end of video recording protocol

Remark:

- Subjects received verbal feedback on performance of the task to ensure a standardized procedure for each movement task during the trial.
- The recorded video tasks were based on DMA-CP principles, and followed the standardized protocol established by Tulloch (Randolph et al., 2005).

Appendix C. Sample of Survey


University of Otago New Zealand | School of Physiotherapy | Centre for Physiotherapy Research | Te Kura Koiri Pai · TePokapu Rakahau Koiri

Exit this survey

Directional Bias Assessment of DMA Clinical Pilates

Demographic data

1. Please enter your personal detail

	Gender	Age	Practicing Country
Personal Detail			

Practicing Country: Other (please specify)

2. Please enter your DMA Clinical Pilates background detail

	Level of training	Was directional bias assessment included?	When was the training?	How often do you use the concept of directional bias assessment since the training?
Detail				

3.
What types of patients do you use DMA concept/ assessment with?
(Select any appropriate)

- ☐ Musculoskeletal
- ☐ Neurological
- ☐ Cardiopulmonary
- ☐ Other (please specify)
- ☐ Paediatric
- ☐ Women's health
- ☐ Chronic Pain

Next

Video 1 of 5

Please answer question 4 & 5 refer to video 1.

Video1

4. Directional bias of subject 1.

	Can you detect a directional bias	What is it?
Video 1		

5. Which following movement(s) helped you to decide the selected bias for subject 1?
(You can indicate more than one)

	Baseline of Single Heel Raise (SHR)	Baseline of Hop	Single Leg Kick (SLK)	Four point kneel (4PK)	Roll ups	SHR after Exercise (Ex.) Intervention	Hop after Ex. Intervention	N/A
Lateral (right/left)	☐	☐	☐	☐	☐	☐	☐	☐
Sagittal (flexion/extension)	☐	☐	☐	☐	☐	☐	☐	☐

Prev Next

Video 2 of 5

Please answer question 6 & 7 refer to video 2.

Video 2

6. Directional bias of subject 2.

	Can you detect a directional bias	What is it?
Video 2		

7. Which following movement(s) helped you to decide the selected bias for subject 2? (You can indicate more than one)

	Baseline of Single Heel Raise (SHR)	Baseline of Hop	Single Leg Kick (SLK)	Four point kneel (4PK)	Roll ups	SHR after Exercise (Ex.) Intervention	Hop after Ex. Intervention	N/A
Lateral (right/left)	☐	☐	☐	☐	☐	☐	☐	☐
Sagittal (flexion/extension)	☐	☐	☐	☐	☐	☐	☐	☐

Prev Next

Video 3 of 5

Please answer question 8 & 9 refer to video 3.

Video 3

8. Directional bias of subject 3.

	Can you detect a directional bias	What is it?
Video 3		

9. Which following movement(s) helped you to decide the selected bias for subject 3? (You can indicate more than one)

	Baseline of Single Heel Raise (SHR)	Baseline of Hop	Single Leg Kick (SLK)	Four point kneel (4PK)	Roll ups	SHR after Exercise (Ex.) Intervention	Hop after Ex. Intervention	N/A
Lateral (right/left)	☐	☐	☐	☐	☐	☐	☐	☐
Sagittal (flexion/extension)	☐	☐	☐	☐	☐	☐	☐	☐

Prev Next

Video 4 of 5

Please answer question 10 & 11 refer to video 4.

Video 4

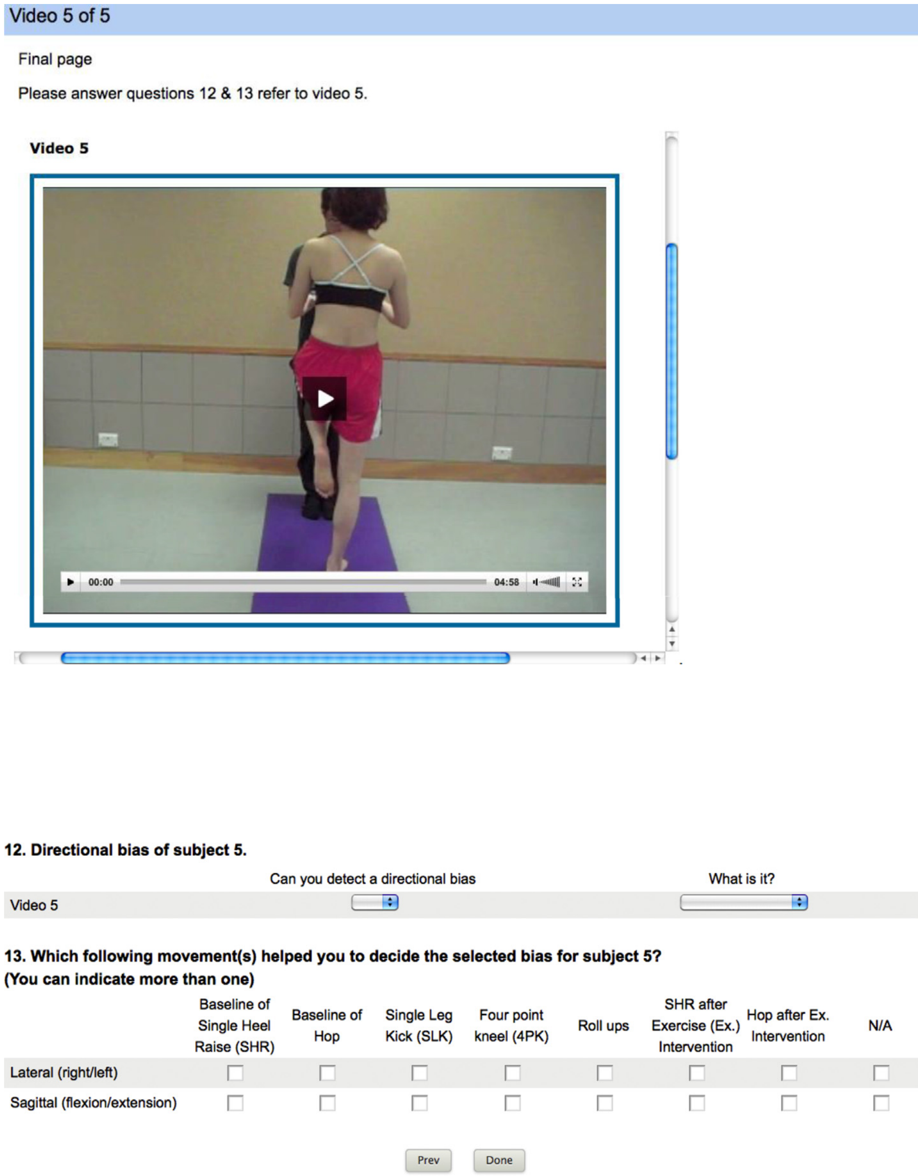
10. Directional bias of subject 4.

	Can you detect a directional bias	What is it?
Video 4		

11. Which following movement(s) helped you to decide the selected bias for subject 4? (You can indicate more than one)

	Baseline of Single Heel Raise (SHR)	Baseline of Hop	Single Leg Kick (SLK)	Four point kneel (4PK)	Roll ups	SHR after Exercise (Ex.) Intervention	Hop after Ex. Intervention	N/A
Lateral (right/left)	☐	☐	☐	☐	☐	☐	☐	☐
Sagittal (flexion/extension)	☐	☐	☐	☐	☐	☐	☐	☐

Prev Next


Video 5 of 5

Final page

Please answer questions 12 & 13 refer to video 5.

Video 5

00:00 04:58

12. Directional bias of subject 5.

	Can you detect a directional bias	What is it?
Video 5		

13. Which following movement(s) helped you to decide the selected bias for subject 5? (You can indicate more than one)

	Baseline of Single Heel Raise (SHR)	Baseline of Hop	Single Leg Kick (SLK)	Four point kneel (4PK)	Roll ups	SHR after Exercise (Ex.) Intervention	Hop after Ex. Intervention	N/A
Lateral (right/left)	☐	☐	☐	☐	☐	☐	☐	☐
Sagittal (flexion/extension)	☐	☐	☐	☐	☐	☐	☐	☐

Prev Done

Appendix D. Statistical Analysis

Free-marginal multirater kappa (K_{free}) (Randolph et al., 2005) was used to measure the overall interrater agreement to detect the presence of a directional bias and classify each individual into 1 of 4 subgroups. The kappa calculations were as follows:

Kappa = [(relative observed agreement among raters – hypothetical probability of chance agreement)/(1-hypothetical probability of chance agreement)] (Randolph et al., 2005).

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Firstly, Let N represent the total number of raters, n the number of ratings (number of videos) per raters, and k the number categories (directional bias of video subjects) into which assignments are made.

The relative observed agreement among raters is set as P_o, the extent to which raters agree for the ith subject need to be calculated, where i is the rater subject, in order to compute how many rater–rater pairs are in agreement relative to the number of all possible rater–rater pairs (Randolph et al., 2005).

$$P_o = \frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij^2} - Nn\right)$$

Secondly, P_e, the hypothetical probability of chance agreement is set equal to $1/k$, where k is the number of rating categories (Randolph et al., 2005).

$$P_e = \frac{1}{k}$$

Finally, free-marginal kappa can be achieved as below: (Randolph et al., 2005)

$$\kappa_{\text{free}} = \frac{\left[\frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij^2} - Nn\right)\right] - \left[\frac{1}{k}\right]}{1 - \left[\frac{1}{k}\right]}$$

References

Cook, C., Heath, F., Thompson, R.L., 2000. A meta-analysis of response rates in Web-or Internet-based surveys. Educ. Psychol. Meas. 60 (6), 821—836.

Cook, C., Massa, L., Harm-Ernandes, I., Segneri, R., Adcock, J., Kennedy, C., Figuers, C., 2007. Interrater reliability and diagnostic accuracy of pelvic girdle pain classification. J. Manip. Physiol. Ther. 30 (4), 252—258.

Dankaerts, W., O'Sullivan, P.B., Straker, L.M., Burnett, A.F., Skouen, J.S., 2006. The inter-examiner reliability of a classification method for non-specific chronic low back pain patients with motor control impairment. Man. Ther. 11 (1), 28—39.

Dettori, J.R., Bullock, S.H., Sutlive, T.G., Franklin, R.J., Patience, T., 1995. The effects of spinal flexion and extension exercises and their associated postures in patients with acute low back pain. Spine 20 (21), 2303—2312.

Eastlack, M.E., Arvidson, J., Snyder-Mackler, L., Danoff, J.V., McGarvey, C.L., 1991. Interrater reliability of videotaped observational gait-analysis assessments. Phys. Ther. 71 (6), 465—472.

Evans, J.R., Mathur, A., 2005. The value of online surveys. Internet Res. 15 (2), 195—219.

Fritz, J.M., Delitto, A., Vignovic, M., Busse, R.G., 2000. Interrater reliability of judgments of the centralization phenomenon and status change during movement testing in patients with low back pain. Arch. Phys. Med. Rehabil. 81 (1), 57—61.

Gladwell, V., Head, S., Haggar, M., Beneke, R., 2006. Does a program of Pilates improve chronic non-specific low back pain? J. Sport Rehabil. 15 (4), 338—350.

Gross, D., Conrad, B., 1991. Issues related to reliability of videotaped observational data. West. J. Nurs. Res. 13 (6), 798—803.

Harris-Hayes, M., Van Dillen, L.R., 2009. The inter-tester reliability of physical therapists classifying low back pain problems based on the movement system impairment classification system. Phys. Med. Rehabil. J. Inj. Funct. Rehabil. 1 (2), 117—126.

Kaplowitz, M.D., Hadlock, T.D., Levine, R., 2004. A comparison of web and mail survey response rates. Public Opin. Q. 68 (1), 94—101.

Karayannis, N.V., Jull, G.A., Hodges, P.W., 2012. Physiotherapy movement based classification approaches to low back pain: comparison of subgroups through review and developer/expert survey. Biomed. Disord. 13 (1), 24—39.

La Touche, R., Escalante, K., Linares, M.T., 2008. Treating non-specific chronic low back pain through the Pilates method. J. Bodyw. Mov. Ther. 12 (4), 364—370.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 159—174.

Lim, E.C., Poh, R.L., Low, A.Y., Wong, W.P., 2011. Effects of Pilates-based exercises on pain and disability in individuals with persistent nonspecific low back pain: a systematic review with meta-analysis. J. Orthop. Sports Phys. Ther. 41 (2), 70—80.

Long, A., Donelson, R., Fung, T., 2004. Does it matter which exercise?: A randomized control trial of exercise for low back pain. Spine 29 (23), 2593—2602.

McKenzie, R., May, S., 2003. The Lumbar Spine: Mechanical Diagnosis and Therapy, vol. 1. Spinal Publications New Zealand, Waikanae, New Zealand.

Randolph, J.J., 2008. Online Kappa Calculator from. http://justus.randolph,name/kappa.

Randolph, J.J., Thanks, A., Bednarik, R., Myller, N., 2005. Free-marginal Multirater Kappa (Multirater *K*-free): an Alternative to Fleiss' Fixed-marginal Multirater Kappa. Citesser.

Rydeard, R., Leger, A., Smith, D., 2006. Pilates-based therapeutic exercise: effect on subjects with nonspecific chronic low back pain and functional disability: a randomized controlled trial. J. Orthop. Sports Phys. Ther. 36 (7), 472—484.

Smits-Engelsman, B.C.M., Fiers, M.J., Henderson, S.E., Henderson, L., 2008. Interrater reliability of the movement assessment battery for children. Phys. Ther. 88 (2), 286—294.

Stolze, L.R., Allison, S.C., Childs, J.D., 2012. Derivation of a preliminary clinical prediction rule for identifying a sub-group of patients with low back pain likely to benefit from Pilates-based exercise. J. Orthop. Sports Phys. Ther. 42 (5), 425—436.

Swaine, B.R., Sullivan, S.J., 1999. Interpreting reliability of early motor function measurement following head injury. Physiother. Theory Pract. 15 (3), 155—164.

Toro, B., Nester, C., Farren, P., 2003. A review of observational gait assessment in clinical practice. Physiother. Theory Pract. 19 (3), 137—149.

Trudelle-Jackson, E., Sarvaiya-Shah, S.A., Wang, S.S., 2008. Interrater reliability of a movement impairment-based classification system for lumbar spine syndromes in patients with chronic low back pain. J. Orthop. Sports Phys. Ther. 38 (6), 371—376.

Tulloch, E., Phillips, C., Sole, G., Carman, A., Abbott, J.H., 2012. DMA clinical Pilates directional-bias assessment: reliability and predictive validity. J. Orthop. Sports Phys. Ther. 42 (8), 676—687.

Vibe Fersum, K., O'Sullivan, P.B., Kvale, A., Skouen, J.S., 2009. Inter-examiner reliability of a classification system for patients with non-specific low back pain. Man. Ther. 14 (5), 555—561.

Xu, X., Chang, C., Faber, G.S., Kingma, I., Dennerlein, J.T., 2011. The validity and interrater reliability of video-based posture observation during asymmetric lifting tasks. Hum. Factors 53 (4), 371—382.