Braz J Phys Ther 29 (2025) 101154

Contents lists available at ScienceDirect

Brazilian Journal of Physical Therapy

journal homepage: www.elsevier.com/locate/bjpt

Original Research

Reliability of the McKenzie Method of Mechanical Diagnosis and Therapy in the examination of spinal pain, including the OTHER classifications Reliability of the McKenzie Method in spinal pain

Hans van Helvoirt [a,*], Henk Tempelman [b], Puck van der Vet [c], Frank van der Vet [d], Job van Helvoirt [e], Richard Rosedale [f], Adri Apeldoorn [g]

[a] Rugpoli Brabant, Tilburg, the Netherlands
[b] Rugpoli Twente, Delden, the Netherlands
[c] VU University, Amsterdam, the Netherlands
[d] Rugpoli Veluwe, Velp, the Netherlands
[e] Technical University of Eindhoven, Eindhoven, the Netherlands
[f] Clinton Physiotherapy, Clinton, Ontario, Canada
[g] Noordwest Ziekenhuisgroep Alkmaar, Rehabilitation Department, Alkmaar, the Netherlands

ARTICLE INFO

Keywords:
Agreement
MDT
OTHER classifications
Reliability

ABSTRACT

Background: The McKenzie Method of Mechanical Diagnosis and Therapy (MDT) is used worldwide to classify and manage musculoskeletal (MSK) problems. The assessment includes a detailed patient history and a specific physical examination. Research has investigated the reliability of the MDT spinal classification system (Derangement syndrome, Dysfunction syndrome, Postural syndrome, and OTHER), however no study has assessed the reliability of the 10 classifications grouped together as OTHER.
Objective: To investigate the inter-rater reliability of MDT trained clinicians when utilising the full breadth of the MDT system for patients with spinal pain.
Methods: Six experienced MDT clinicians each submitted potentially eligible MDT assessment forms of 30 consecutive patients. A MSK physician and a faculty of the McKenzie Institute checked the 180 forms for eligibility and completeness, where a provisional MDT classification was blinded. Apart from their own assessment forms, the six MDT clinicians each classified 150 forms. Each patient could be classified into 1 of 13 diagnostic classifications (Derangement syndrome, Dysfunction syndrome, Postural syndrome, and 10 classifications grouped as OTHER). Reliability was determined using Fleiss' Kappa (k).
Results: The reliability among six MDT clinicians classifying 150 patient assessment forms was almost perfect (Fleiss' $\kappa = 0.82$ [95% CI 0.80, 0.85]).
Conclusions: Among experienced MDT clinicians, the reliability in classifying patient assessment forms of patients with spinal pain is almost perfect when the full breadth of the MDT system is used. Future research should investigate the reliability of the full breadth of the MDT system among clinicians with lower levels of training.

Introduction

Mechanical Diagnosis and Therapy (MDT), also called the "McKenzie Method" has been widely used by physical therapists and other health care practitioners as an approach for patients with musculoskeletal disorders.[1,2] The MDT assessment includes a detailed patient history and a specific physical examination, within a biopsychosocial context, recognising potential drivers of pain and disability.[2-6] The physical examination includes postural observation, movement loss, neurological testing, the establishing and the retesting of baselines, and the use of repeated movements, sustained positions, and other testing (e.g., sacroiliac joint pain provocation tests). The symptomatic and mechanical responses to different loading strategies guides the clinician towards a provisional MDT classification and classification-based treatment. In consecutive sessions, initial provisional classifications can be confirmed, rejected, or modified.[7,8] Apart from specific exercises and postures, MDT aims to promote patient-specific education and self-management, embedded in a biopsychosocial context.[3,4,9]

* Corresponding author at: Scheepswerf 2, 5256, PL, Heusden, the Netherlands.
E-mail address: hvhmck@gmail.com (H. van Helvoirt).

https://doi.org/10.1016/j.bjpt.2024.101154
Received 31 July 2023; Received in revised form 13 April 2024; Accepted 13 November 2024
Available online 13 December 2024

Originally, the MDT system consisted of three main MDT syndromes to classify patients with musculoskeletal (MSK) conditions, that is Derangement, Dysfunction, and Postural. Prevalence data over the last two decades have shown that the Derangement syndrome has a high prevalence in the spine while Dysfunction and Postural syndromes are relatively rare.[10] People not exhibiting the characteristics of one of these three syndromes were classified as 'OTHER'. Currently, OTHER contains 10 specific classifications.[10] Examples include Spinal Stenosis, Mechanically Inconclusive, and Trauma. The 13 classifications (Derangement, Dysfunction, and Postural syndrome and the 10 classifications grouped as OTHER) with their own operational definitions are described in Supplementary material 1 and 2.

In MDT, each classification is matched with a specific intervention and therefore reliability is key for the selection of appropriate management, which ultimately determines the treatment outcome. If there is unacceptable reliability, management following the classification may be inappropriate, as it may be based on an incorrect classification. Several studies have been performed on the reliability of the three McKenzie syndromes and the OTHER category as a whole grouping.[11-16]

Kilpikoski et al. found substantial reliability when patients with low back pain were classified into the McKenzie main syndromes ($k = 0.6$) and agreement was 95% between two trained MDT therapists.[13] Clare *et al.* showed substantial to almost perfect reliability for syndrome classification ($k = 0.84$) with 96% agreement for the total patient pool, ($k = 1.0$) with 100% agreement for lumbar patients, and ($k = 0.63$) 92% agreement for cervical patients between two trained therapist.[11] Razmjou et al. indicated good inter-examiner reliability between two therapists who were trained in MDT ($k = 0.7$).[14] Dionne et al. found moderate reliability for diagnosis [κ=0.55, $P < 0.001$, confidence intervals (CI) 0.52, 0.58], for raters with different levels of MDT training in 20 videotaped cervical patients.[17] Yet, to our knowledge, no studies have investigated the inter-rater reliability of clinicians using the 10 classifications with specific diagnostic criteria grouped under OTHER. One of the difficulties has been the low prevalence rate of each of those 10 other diagnostic classifications in primary care. For example, a prevalence survey by May and Rosedale based on 750 patients from 54 therapists from 15 different countries who worked in a variety of healthcare settings showed only 23% of the patients being classified as one of the 10 classifications grouped as OTHER.[8]

In Rugpoli, a Dutch medical centre that provides secondary and tertiary level care, patients are frequently classified as one of the OTHER classifications. Analysis of data for 3798 patients from Rugpoli during 2019, from eight physical therapists, showed that 63% of the patients were classified as one of the OTHER classifications. Table 1 shows the prevalence rates of the three main MDT syndromes and the 10 OTHER classifications at the Rugpoli facility. Due to the high prevalence rates of the classifications within the OTHER group in comparison with primary care, the Rugpoli facility is an ideal setting for conducting a reliability study focusing on these subgroups. Thus, the purpose of this study was to examine the reliability between MDT clinicians of the full breadth of the MDT syndrome classification for patients with spinal pain using MDT assessment forms.

Table 1
Prevalence of the three main Mechanical Diagnosis and Therapy (MDT) syndromes and the 10 classifications under OTHER in Rugpoli in 2019 ($n = 3798$).

	Main MDT syndromes	n (%)
1	Derangement	1253 (33)
2	Dysfunction	151 (4)
3	Postural	0 (0)
	10 classification under OTHER	
4	Mechanically Unresponsive Radicular Syndrome (MUR)	911 (24)
5	Mechanically Inconclusive	833 (22)
6	Spinal Stenosis	265 (7)
7	Sacroiliac Joint (SIJ) Pain	76 (2)
8	Inflammatory Arthropathy/Chemical	151 (4)
9	Chronic Pain Syndrome	151 (4)
10	Structurally Compromised	0 (0)
11	Trauma/Recovering Trauma	0 (0)
12	Serious Pathology	0 (0)
13	Post Surgery	10 (0.003)

Methods

Assessment forms of real patients were used in this study to examine the reliability of MDT classification. Patient assessment forms are often used to examine reliability.[18] Correctly designed, patient assessment forms are practical to use, avoid ethical problems, are inexpensive, and can collect data from different sources.[18-22] Ethical approval for the study was obtained from the Health Science Research Ethics Board at the VU University of Amsterdam, the Netherlands, in March 2020.

Study design

A two-phase vignette-based reliability study about consecutive patient files of real patients.

A two-phase study was conducted. For the first phase, five diplomate MDT therapists and one credentialed MDT therapist from different Rugpoli centres were requested to each generate 30 patient assessment forms of consecutive patients with spinal pain from Rugpoli intakes in 2019. This resulted in a total of 180 patient assessment forms. All patient assessment forms were electronically sent to the first author HH and FV, two authors of this paper, who checked the assessment forms for completeness, MDT classification characteristics, ambiguity regarding clinical presentation, a provisional MDT classification, and ethical board rules (such as the absence of personal data and age indicated only in decades, etc.). Rugpoli uses an assessment form similar to the MDT assessment form but written in a different format that aligns with the clinic's information technology system. Conclusions and any documentation of provisional MDT classification or management were removed from the patient files. In case of uncertainties or discrepancies, clinicians were consulted to discuss and confirm the accuracy of the case. 29 of 180 patient assessment forms were discussed by HH and FV with the senders, primarily because neurological examination findings were not recorded in the classification of Mechanically Unresponsive Radicular Syndrome. This non-recording was due to the setting of the clinics where a neurologist is the first examiner of a patient with radiculopathies. These neurology notes could be seen by the MDT therapist and creator of the assessment form but were not always transcribed to the assessment form. Each assessment form was randomly assigned a number from 1 to 180 to enable tracking of responses and data collection.

In the second phase, these six MDT clinicians rated all the clinical assessments forms except the 30 they had personally created. Thus, each rater examined a total of 150 assessment forms and each form was rated by five of the six raters. They were instructed to review each patient assessment forms based on its history and clinical presentation and to assign an MDT classification syndrome (Derangement, Dysfunction, or Postural) or one of the 10 OTHER classifications, including Serious Pathology, Chronic Pain Syndrome, Mechanical Inconclusive, Mechanically Unresponsive Radiculopathy (MUR), Inflammatory Arthropathy (including active Modic signs), Sacroiliac Joint Pain, Spinal Stenosis, Trauma, Post Surgery, and Structurally Compromised. All raters were blinded to the provisional MDT classification originally assigned to the patient assessment forms by its creator in phase one and to the classification of other raters. Raters were not aware of the figures in Table 1. Informed consent was obtained from each clinician.

Characteristics of the researchers and the raters

The study was performed by two researchers and six raters. One of the two researchers (HH) is a Senior Faculty of the McKenzie Institute and has 32 years of clinical experience as an MDT Diploma physical therapist working with patients with musculoskeletal spinal pain. The

Table 2
Characteristics of the MDT therapists that rated the assessments forms.

Rater	Age (yrs)	PT (yrs)	Credentialed MDT (yrs)	Diploma MDT (yrs)	Rugpoli (yrs)
1	51	28	24	20	8
2	41	19	12	0	10
3	54	31	24	20	11
4	47	23	18	7	14
5	48	25	17	7	12
6	32	11	10	3	6

yrs = years.
note: Rater two had basic training in MDT (Credentialed) and further in-company training by two faculty of the McKenzie Institute working in Rugpoli, but no formal diploma MDT training.

second researcher (FV) is a musculoskeletal physician with 30 years of clinical experience. The characteristics of the 6 raters are displayed in Table 2.

Sample size

For the sample size calculation with multiple outcomes and raters, the method described by Rotondi and Donner was used.[23,24] Based on the levels of reliability in earlier MDT studies, kappa was set at 0.8 (0.7 lower limit, 0.9 upper limit).[11,13,14] With an alpha of 0.05 for six raters (phase two) and using the prevalence rates of the three MDT syndromes and the OTHER categories shown in Table 1 as outlined by May and Rosedale, 150 assessment forms were required.[8] The sample size was calculated using the R Project for Statistical Computing (Vienna, Austria).

Analysis

To determine interrater reliability between the six raters, Fleiss' κ and the 95% confidence interval (CI) were calculated with the provisional MDT diagnosis as the basis. Also, Fleiss' κ values, 95% CIs and agreement values were calculated for each diagnostic classification.[25-27] Besides, the κ statistic and percentage agreement were analysed and reported between each pair of raters.[28] Overall κ for three or more raters may result in a better representation of reliability, however it may mask extreme cases of agreement or disagreement among paired raters.[29] Reliability and agreement measures provide different types of information. For the present study reliability is most appropriate, because it provides information about the ability of the MDT classification system to distinguish between cases.[30] For interpretations of the κ values the guidelines of Landis and Koch were used: 0.01 to 0.20 slight, 0.21 to 0.40 fair, 0.41 to 0.60 moderate, 0.61 to 0.80 substantial, 0.81 to 1.00 almost perfect agreement.[31] The data were analyzed using IBM SPSS Statistics Version 29.0.

Results

The 180 patient assessment forms consisted of 157 lumbar cases (109 with symptoms in leg(s) (below buttocks)), 20 cervical cases (17 with symptoms in arms (below shoulder), and 3 thoracic cases (all with

Table 3
Agreement (percentage) and reliability (κ values) between each pair of 6 raters.

Kappa values (% agreement)					
Rater	1	2	3	4	5
1					
2	0.86 (89)				
3	0.84 (88)	0.84 (88)			
4	0.85 (88)	0.84 (88)	0.86 (88)		
5	0.82 (87)	0.82 (87)	0.86 (89)	0.84 (88)	
6	0.77 (83)	0.79 (84)	0.80 (85)	0.79 (84)	0.77 (83)

Table 4
Prevalence of the provisional diagnostic Mechanical Diagnosis and Therapy (MDT) classification from the 180 patient assessment forms, the classifications by the six raters and Fleiss' kappa values for each classification.

Classification	Provisional diagnostic classification of the 180 assessment forms, n (%)	Six raters classified each 150 patient assessment forms (6 × 150 = 900 classifications)											Fleiss' kappa (95%CI)	Agreement, %
		Der	Dys	Post	MUR	MI	Stenosis	SIJ	Infl	Chr Pain	Struc Com	Trauma	Ser Path	Post Surg		
Main MDT syndromes																
1	Der.	67 (37)	327				7		1						0.94 (0,89-0.99)	**98**
2	Dys.	5 (3)		18			7								0.72 (0.67-0.76)	**72**
3	Post.	0 (0)													NA	**NA**
10 subgroups under the classification OTHER																
4	M.U.R.	43 (24)				193	21	1							0.88 (0.83-0.92)	**90**
5	M.I.	34 (19)	4			4	152	2	3	3	2				0.71 (0.66-0.75)	**89**
6	Spinal Stenosis	10 (6)	1			8	6	33			2				0.58 (0.53-0.63)	**66**
7	SIJ	7 (4)	1			2	8		24						0.65 (0.60-0.70)	**69**
8	Infl.	8 (4)					2			38					0.90 (0.85-0.95)	**95**
9	Chr. pain	6 (3)					4			1	25				0.75 (0.70-0.80)	**83**
10	Struc. com.	0													NA	**NA**
11	Trauma	0													NA	**NA**
12	Ser. Path.	0													NA	**NA**
13	Post. Surg.	0													NA	**NA**

Der, Derangement syndrome; Dys, Dysfunction syndrome ; Post, Posture syndrome; MUR, Mechanically Unresponsive Radicular syndrome; MI, Mechanically Inconclusive; SIJ, sacroiliac joint pain; infl, Inflammatory Arthropathy / Chemical; Chr Pain, Chronic Pain syndrome; Struc Com, Structurally Compromised; Trauma, Trauma/ Recovering Trauma; Ser Path, Serious Pathology; Post Surg, Post Surgery; CI, confidence interval; NA, not applicableTable 4 shows the paired comparison of reliability and agreement in patient assessment forms among the six raters. κ values ranged from 0.83 to 0.89 and agreement values from 83% to 89%.

symptoms referring into rib regions), patients were 18 years or older). All six raters rated 150 assessment forms each and there were no missing data. Fleiss' κ was run to determine inter-rater reliability between raters and was found to be 0.82 (95% CI 0.80, 0.85), indicating almost perfect reliability.

Paired comparison of agreement and reliability in patient assessment forms among the six raters. Agreement values ranged from 83% to 89% and κ values from 0.77 to 0.86 (Table 3).

The Derangement syndrome classification had the highest level of interrater reliability (Fleiss' κ = 0.94, 95% CI 0.89, 0.99), while the Spinal Stenosis classification had the lowest level of reliability (Fleiss' κ = 0.58, CI 0.53, 0.63). Agreement values varied between 98% (Derangement syndrome) and 66% (Spinal Stenosis) (Table 4). The Fleiss' k values indicate that the level of reliability for individual classifications varies between almost perfect (Derangement syndrome) and moderate (Spinal Stenosis).

Chr Pain, Chronic Pain syndrome; CI, confidence interval; Der, Derangement syndrome; Dys, Dysfunction syndrome; infl, Inflammatory Arthropathy / Chemical; MI, Mechanically Inconclusive; MUR, Mechanically Unresponsive Radicular syndrome; NA, not applicable; Post, Posture syndrome; Post Surg, Post Surgery; Ser Path, Serious Pathology; SIJ, sacroiliac joint pain; Struc Com, Structurally Compromised; Trauma, Trauma/ Recovering Trauma

Discussion

This is the first study that assessed the inter-rater reliability of the McKenzie Method of MDT in the examination of spinal pain that has encompassed all the individual OTHER classifications. This study found almost perfect inter-rater reliability between the judgements of six raters using patient assessment forms. Additionally, high κ values were found between each pair of raters, indicating that there was also almost perfect agreement between each pair of raters.

The results of the current study are in concordance with the conclusions of the review of Garcia *et al.*[32] They concluded that the MDT system demonstrates acceptable inter-rater reliability when classifying patients with back pain into main syndromes by therapists who have completed the credentialing examination. Werneke et al. reported that lumbar classification among therapists with pre-credentialed level of training only had fair to moderate reliability (κ = 0.37 to 0.44), despite high observed agreement (86–91%).[33] However, it has been proposed that this paradox may be due to the sensitivity of the kappa statistic to skewed prevalence of ratings, and that drawing conclusions based solely on the kappa statistic may be misleading.[28,34,35] The high prevalence of Derangement (81–86%) and the relatively low prevalence of other classifications (0–4.6%) in this study likely contribute to the low kappa values, despite the high observed agreement.[28,29,33-36] In the current study the prevalence rates of all the classifications were much less skewed and therefore observed agreement and κ values are in balance.

The highest level of reliability was observed for Derangement syndrome (κ = 0.94), while the lowest level of reliability was found for Spinal Stenosis (κ = 0.58). The difference in reliability between these diagnostic classifications could be attributed to their respective clinical presentations. Derangement syndrome is characterized by a sustained reduction or elimination of symptoms through repetitive or sustained loading strategies in a specific direction, making it easily identifiable.[3,10,37] On the other hand, the lower level of agreement for Spinal Stenosis could be due to the overlapping aspects in the operational definitions with MUR. The rest of the OTHER classifications achieved substantial to almost perfect levels of reliability, indicating that the operational definitions for different OTHER classifications are well-described and recognized during initial visits. (See supplementary material 1 and 2, from Part A Course Manual, McKenzie Institute International ©)

One of the OTHER classifications is Serious Pathology. Any presentation of Serious Pathology is normally identified by the medical doctor (MSK physician or neurologist) who initially screens the patients, being the first in line during the intake procedure in Rugpoli, and therefore this classification was not seen by the creators of the patient assessment forms and raters of this study.

Strengths and limitations

To strengthen the study, we used assessment forms of real patients and a thorough power calculation resulting in 150 patients rated by six reviewers. One of the limitations was that raters where highly trained and experienced in utilising the MDT Method in a clinical setting where prevalence rates of OTHER classifications may differ substantially from primary care. This makes these data less generalisable to other, less trained physical therapist or those working in a clinical setting where the prevalence rates for OTHER classifications are low. Also, using written patient assessment forms eliminates the potential error created by a change in the patient presentations between raters but may not represent the full clinical picture as seen in practice, where nonverbal communication possibly influences decisions in testing and clinical reasoning. This can make classification easier and potentially has a positive impact on kappa values. Further research with real patients still needs to be done to establish full reliability.

Conclusion

This is the first study to our knowledge that assessed the inter-rater reliability of the McKenzie Method of MDT in the examination of spinal pain that has encompassed all the individual OTHER classifications. This study found almost perfect inter-rater reliability between the judgements of six raters using patient assessment forms. The level of reliability for individual classifications varied between substantial to almost perfect. Additionally, almost perfect inter-rater reliability values were found between each pair of raters, indicating that there was also almost perfect agreement between each pair of raters

Declaration of competing interest

The researchers declare no competing interest.

Acknowledgements

Henk Tempelman, Hanneke Meihuizen, Daniel Poynter, Peter Hagel, Marijke Mol, Loes Geerlings, for creating and rating patient assessment forms.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.bjpt.2024.101154.

References

1. Billis EV, McCarthy CJ, Oldham JA. Subclassification of low back pain: a cross-country comparison. *Eur Spine J*. 2007;16(7):865–879. https://doi.org/10.1007/s00586-007-0313-2.
2. Stynes S, Konstantinou K, Dunn K. Classification of patients with low back-related leg pain: a systematic review. *BMC Musculoskeletal Dis*. 2016;17:226. https://doi.org/10.1186/s12891-016-1074-z.
3. McKenzie RA, May S. *The Lumbar Spine: Mechanical Diagnosis & Therapy*. 2nd ed. New Zealand: Spinal Publications; 2003.
4. McKenzie RA, May S. *The Cervical Spine: Mechanical Diagnosis & Therapy*. 2nd ed. New Zealand: Spinal Publications; 2006.
5. Werneke MW, Hart DL, George SZ, Deutcher D, Stratford PW. Change in psychosocial distress associated with pain and functional status outcomes in patients with lumbar impairments referred to physical therapy services. *J Orthop Sports Phys Ther*. 2012;41(12):969–980. https://doi.org/10.2519/jospt.2011.3814.
6. Werneke M, Edmond S, Deutscher D, et al. Effect of adding McKenzie Syndrome, centralization, directional preference, and psychosocial classification variables to a risk adjusted model predicting functional status outcomes for patients with lumbar

impairments. *J Orthop Sports Phys Ther*. 2016;46(9):726–741. https://doi.org/10.2519/jospt.2016.6266.
7. Werneke MW, Hart DL, Cook D. A descriptive Study of the Centralization Phenomenon. *Spine (Phila Pa 1976)*. 1999;24(7):676–683. https://doi.org/10.1097/00007632-199904010-00012.
8. van Helvoirt H, Apeldoorn A, Knol D, et al. Transforaminal epidural steroid injections influence Mechanical Diagnosis and Therapy (MDT) pain response classification in candidates for lumbar herniated disc surgery. *J Back Musculoskelet Rehabil*. 2016;29(2):351–359. https://doi.org/10.3233/BMR-160662.
9. Kuhnow A, Kuhnow J, Ham D, Rosedale R. The McKenzie Method and its association with psychosocial outcomes in low back pain: a systematic review. *Physiotheraoy Theory and Practice*. 2021;37(12):1283–1297. https://doi.org/10.1080/09593985.2019.1710881.
10. May S, Rosedale R. An international survey of the comprehensiveness of the McKenzie classification system and the proportions of classifications and directional preferences in patients with spinal pain. *Musculoskelet Sci Pract*. 2019;39:10–15. https://doi.org/10.1016/j.msksp.2018.06.006.
11. Clare HA, Adams R, Maher CG. Reliability of McKenzie classifications of patients with cervical or lumbar pain. *J Manip Physiol Ther*. 2005;28(2):122–127. https://doi.org/10.1016/j.jmpt.2005.01.003.
12. Fritz JM, Delitto A, Vignovic M, Busse RG. Interrater reliability of judgments of the centralization phenomenon and status change during movement testing in patients with low back pain. *Arch Phys Med Rehabil*. 2000;81(1):57–61. https://doi.org/10.1016/S0003-9993(00)90222-3.
13. Kilpikoski S, Airaksinen O, Kankaanpaa M, Leminen P, Videman T, Alen M. Interexaminer reliability of low back pain assessment using the McKenzie method. *Spine (Phila Pa 1976)*. 2002;8:E207–E214. https://doi.org/10.1097/00007632-200204150-00016.
14. Razmjou H, Kramer JF, Yamada R. Intertester reliability of the McKenzie evaluation in assessing patients with mechanical low-back pain. *J Orthop Sports Phys Ther*. 2000;30(7):368–383. https://doi.org/10.2519/jospt.2000.30.7.368.
15. Werneke MW, D.L.Hart D Deutscher. Clinician's ability to identify neck and low back interventions: an inter-rater chance-corrected agreement pilot study. *Spine (Phila Pa 1976)*. 2013;39:172–181. https://doi.org/10.1179/2042618611Y.0000000001.
16. Gutke A, Kjellby-Wendt G, Oberg B. The inter-rater reliability of a standardised classification system for pregnancy-related lumbopelvic pain. *Man Ther*. 2010;15(1):13–18. https://doi.org/10.1016/j.math.2009.05.005.
17. Dionne CP, Bybee RF, Tomaka J. Inter-rater reliability of McKenzie assessment in patients with neck pain. *Physiotherapy*. 2006;92(2):75–82. https://doi.org/10.1016/j.physio.2005.12.003.
18. S.C.Evans MCRoberts, Keeley JW, Blossom JB, et al. Vignette methodologies for studying clinicians' decision-making: validity, utility, and application in ICD-11 field studies. *Int J Clin Health Psychol*. 2015;15(2):160–170. https://doi.org/10.1016/j.ijchp.2014.12.001.
19. Peabody JW, Luck J, Glassman P, Dresselhaus TR, Lee M. Measuring the quality of physician practice by using clinical vignettes: a prospective validation study. *Ann Intern Med*. 2004;141(10):771–780. https://doi.org/10.1001/jama.283.13.1715.
20. Gould D. Using vignettes to collect data for nursing research studies: how valid are the findings? *J Clin Nurs*. 1996;5(4):207–219. https://doi.org/10.1111/j.1365-2702.1996.tb00253.x.
21. Veloski J, Tai S, Evans AS, Nash DB. Clinical vignette-based surveys: a tool for assessing physician practice variation. *Am J Med Qual*. 2005;20(3):151–157. https://doi.org/10.1177/1062860605274520.
22. Rutten GM, Harting J, Rutten ST, Bekkering GE, Kremers SP. Measuring physiotherapists' guideline adherence by means of clinical vignettes: a validation study. *J Eval Clin Pract*. 2006;12(5):491–500. https://doi.org/10.1111/j.1365-2753.2006.00699.x.
23. Rotondi MA, Donner A. A confidence interval approach to sample size estimation for interobserver agreement studies with multiple raters and outcomes. *J Clin Epidemiol*. 2012;65(7):778–784. https://doi.org/10.1016/j.jclinepi.2011.10.019.
24. Donner A, Rotondi MA MA. Sample size estimation functions for studies of interobserver agreement. *Int J Biostat*. 2010;6(1). https://doi.org/10.2202/1557-4679.1275.
25. Fleiss JI. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76(5):378–382. https://doi.org/10.1037/h0031619.
26. Fleiss JI, Nee JC, Landis JR. Large sample variance of kappa in the case of different sets of raters. *Psychol Bull*. 1979;86(5):974–977. https://doi.org/10.1177/0013164420973080.
27. G. Cardillo. Fleiss'es kappa: compute the Fleiss'es kappa for multiple raters. https://nl.mathworks.com/matlabcentral/fileexchange/15426-fleiss.
28. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol*. 1993;46(5):423–429. https://doi.org/10.1016/0895-4356(93)90018-V.
29. O'leary S, Lund M, Ytre-Hauge TJ, Naess K, Nagelstad Dalland L, McPhail SM. Pitfalls in the use of kappa when interpreting agreement between multiple raters in reliability studies. *Physiotherapy*. 2014;100(1):27–35. https://doi.org/10.1016/j.physio.2013.08.002.
30. Kottner J, Streiner L. The difference between reliability and agreement. *J Clin Epidemiol*. 2011;64(6):701–702. https://doi.org/10.1016/j.jclinepi.2010.12.001.
31. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics*. 1977;33(1):159–174. https://doi.org/10.2307/2529310.
32. Garcia AN, Menezes C, de Souza FS. Reliability of Mechanical Diagnosis and Therapy system in patients with spinal pain: a systematic review. *J Orthop Sports Phys Ther*. 2018;48(12):923–933. https://doi.org/10.2519/jospt.2018.7876.
33. Werneke MW, Deutscher D, Hart D, et al. McKenzie lumbar classification: inter-rater agreement by physical therapists with different levels of formal McKenzie postgraduate training. *Spine (Phila Pa 1976)*. 2014;39(3):182–190. https://doi.org/10.1097/BRS.0000000000000117.
34. Zapf A, Castell S, Morawietz L, Krach A. Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Mediacal Research Methodology*. 2016;16:93. https://doi.org/10.1186/s12874-016-0200-9.
35. de Vet HC, Mokkink LB, Terwee CB, Hoekstra OS, Knol DL. Clinicians are right not to like Cohen's κ. *BMJ*. 2013;346:1–7. https://doi.org/10.1136/bmj.f2125.
36. McKenzie DP, Mackinnon AJ, Péladeau N, et al. Comparing correlated kappas by resampling: is one level of agreement significantly different from another? *J Psychiatr Res*. 1996;30(6):83–492. https://doi.org/10.1016/s0022-3956(96)00033-7.
37. May S, Aina A. Centralization and directional preference: a systematic review. *Man Ther*. 2012;17(6):497–506. https://doi.org/10.1016/j.math.2012.05.003.