Interrater and Intrarater Reliability of the Beighton Score

A Systematic Review

Lauren N. Bockhorn,* MD, Angelina M. Vera,* MD, David Dong,* BS, Domenica A. Delgado,* MBA, Kevin E. Varner,* MD, and Joshua D. Harris,*† MD

Investigation performed at Houston Methodist Orthopedics and Sports Medicine, Houston, Texas, USA

Background: The Beighton score is commonly used to assess the degree of hypermobility in patients with hypermobility spectrum disorder. Since proper diagnosis and treatment in this challenging patient population require valid, reliable, and responsive clinical assessments such as the Beighton score, studies must properly evaluate efficacy and effectiveness.

Purpose: To succinctly present a systematic review to determine the inter- and intrarater reliability of the Beighton score and the methodological quality of all analyzed studies for use in clinical applications.

Study Design: Systematic review; Level of evidence, 3.

Methods: A systematic review of the MEDLINE, Embase, CINAHL, and SPORTDiscus databases was performed. Studies that measured inter- or intrarater reliability of the Beighton score in humans with and without hypermobility were included. Non-English, animal, cadaveric, level 5 evidence, and studies utilizing the Beighton score self-assessment version were excluded. Data were extracted to compare scoring methods, population characteristics, and measurements of inter- and intrarater reliability. Risk of bias was assessed with the COSMIN (Consensus-Based Standards for the Selection of Health Measurement Instruments) 2017 checklist.

Results: Twenty-four studies were analyzed (1333 patients; mean ± SD age, 28.19 ± 17.34 years [range, 4-71 years]; 640 females, 594 males, 273 unknown sex). Of the 24 studies, 18 reported raters were health care professionals or health care professional students. For interrater reliability, 5 of 8 (62.5%) intraclass correlation coefficients and 12 of 19 (63.2%) kappa values were substantial to almost perfect. Intrarater reliability was reported as excellent in all studies utilizing intraclass correlation coefficients, and 3 of the 7 articles using kappa values reported almost perfect values. Utilizing the COSMIN criteria, we determined that 1 study met "very good" criteria, 7 met "adequate," 15 met "doubtful," and 1 met "inadequate" for overall risk of bias in the reliability domain.

Conclusion: The Beighton score is a highly reliable clinical tool that shows substantial to excellent inter- and intrarater reliability when used by raters of variable backgrounds and experience levels. While individual components of risk of bias among studies demonstrated large discrepancy, most of the items were adequate to very good.

Keywords: Beighton score; hypermobility; systematic review; interrater; intrarater

The Beighton score is the cornerstone for diagnosing hypermobility syndromes, including hypermobility spectrum disorder or hypermobile Ehlers-Danlos syndrome.[13,59] The original criteria do not provide a detailed description,[6] which leaves them open for interpretation and uncertainty of application. No threshold score is determined by the original description,[6] nor is there consensus throughout the literature on what defines hypermobility.[24,34] However, variations are seen in hypermobility depending on age, sex, and race; thus, some experts believe that thresholds should be individualized to subpopulations.[51,52] Given the imprecision of the Beighton score, studies utilizing it may be inconsistent in starting positions, performance, and benchmarks.[34] Questions left unanswered by the Beighton score include whether the tests should be performed actively by the respondent or passively by the clinician and whether a warm-up period is required.[35] The risk of these inherent shortcomings is that a lack of specificity could affect the score's generalizable applicability and reliability. In addition, the Beighton score does not account for symptoms. Laxity is defined as excessive motion in a specific joint in an asymptomatic individual. "Excessive" relative to a joint,

The Orthopaedic Journal of Sports Medicine, 9(1), 2325967120968099
DOI: 10.1177/2325967120968099

is defined as abnormally increased or supraphysiologic motion, also known as "hypermobility." "Instability" is defined as excessive motion in a specific joint in a symptomatic individual. The key distinction between laxity and instability is the absence (former) or presence (latter) of symptoms.

Historically, studies have consistently reported excellent reliability of the Beighton score. However, recent systematic reviews have reported these studies to show conflicting evidence, and they have cited concerns with the methodology based on requirements with COSMIN (Consensus-Based Standards for the Selection of Health Measurement Instruments) criteria that are clinically inapplicable to this score.[17,36] The training and experience of raters[26,42] and the time between examinations[33] have the potential to affect the measures of Beighton score reliability according to the current COSMIN criteria. Reliable, accurate, and precise measures for hypermobility are necessary for operative and nonoperative musculoskeletal care for clinicians and surgeons. Specifically, they can guide treatment choices in patellofemoral,[10] shoulder,[53] and hip instability[46] as well as anterior cruciate ligament (ACL) reconstruction.[41,55]

Owing to the significant heterogeneity in evidence regarding the Beighton score, the purpose of this investigation was to succinctly present a systematic review to determine the inter- and intrarater reliability of the Beighton score and the methodological quality of all analyzed studies in the context of clinical applicability. We hypothesized that this systematic review will demonstrate excellent inter- and intrarater reliability and substantial methodological quality that is satisfactory for surgeons' clinical use.

METHODS

The review protocol was registered via the National Institute for Health Research's PROSPERO International Prospective Register of Systematic Reviews (CRD42018081703).[28] The systematic review was conducted according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines.[43] Utilizing PICO (population, intervention, comparison, outcome) to fit a measurement tool, we examined research addressing humans of any age, degree of hypermobility, the Beighton score, and inter- and intrarater reliability. Therefore, it was determined that studies evaluating the clinical Beighton score between and among raters as a primary or secondary outcome would be included and all others would be considered the wrong outcome. Studies that utilized the Beighton self-assessment exclusively, in which patients independently measured and reported their own score, were excluded. Reviews, abstracts, theses, unpublished studies, articles not available in English, and studies with animal or cadaveric subjects were also excluded.

A systematic computerized search (Appendix 1) was conducted by 1 author (L.N.B.) on January 30, 2018, in 4 databases (MEDLINE, Embase, CINAHL, and SPORTDiscus) with no limitations on dates of inclusion. To reduce the search bias, the search strategy was conducted using Medical Subject Headings. A search in ClinicalTrials.gov was also conducted to identify any possible ongoing studies. The search terms included, but were not limited to the following: Beighton, joint laxity, hypermobility, reproducibility of results, observer variation, reliability, interrater, or intrarater (Appendix 1).

Identified records were imported to the systematic review software Rayyan (Qatar Computing Research Institute),[48] and duplicates were removed. Articles were screened in a 2-step process, first by title and abstract according to exclusion criteria. Second, articles included by abstract were imported into Rayyan; full texts were made available; and 2 authors (L.N.B. and A.M.V.) independently screened by reading the article abstract and the article full text for inclusion according to both eligibility criteria. Disagreements concerning final inclusion were settled by consensus between these authors during a deliberation session.

The data extraction sheet was developed according to the Cochrane Consumers and Communication Review Group's data extraction template,[30] was pilot tested on 3 randomly selected included studies, and then refined accordingly. One review author (L.N.B.) extracted the data from included studies, which the second author (A.M.V.) verified. Disagreements were resolved by discussion between them; if no agreement could be reached, it was planned that a third author (J.D.H.) would decide. No authors were contacted for additional information, and all missing data were labeled "not specified."

The included articles were independently assessed by 2 authors (L.N.B. and A.M.V.) for risk of bias using the COSMIN checklist.[44] The complete COSMIN checklist includes 12 boxes, covering internal consistency, reliability, measurement error, validity, and responsiveness. This review exclusively evaluated reliability (COSMIN box 6), which was determined to be crucial to the context in which inter- and intraobserver values were interpreted. The overall methodological quality of a study is determined by the lowest rating among the items in the reliability box (ie, "the worst score counts" principle), including "very good,"

†Address correspondence to Joshua D. Harris, MD, Houston Methodist Orthopedics and Sports Medicine, 6445 Main Street, Suite 2500, Houston, TX 77030, USA (email: joshuaharrismd@gmail.com).

*Houston Methodist Orthopedics and Sports Medicine, Houston, Texas, USA.

Final revision submitted May 27, 2020; accepted June 23, 2020.

One or more of the authors has declared the following potential conflict of interest or source of funding: A.M.V. has received educational support from Arthrex/Medinc and DePuy. K.E.V. has received consulting fees from DePuy, In2Bones, and Wright Medical and receives royalties from and has stock/stock options in In2Bones and Wright Medical. J.D.H. has received research support from Arthrex/Medinc, DePuy, and Smith & Nephew; consulting fees from NIA Magellan and Smith & Nephew; speaking fees from Ossur and Smith & Nephew; and royalties from SLACK. AOSSM checks author disclosures against the Open Payments Database (OPD). AOSSM has not conducted an independent investigation on the OPD and disclaims any liability or responsibility relating thereto.

TABLE 1
Extracted Data[a]

Population Description	Test Conditions	Whether Test Conditions Were Similar for the Measurements
Number of participants	Beighton score modifications	Participant sequence generation
Age	Examination setting	Whether sequence of participants was concealed
Sex	Number of raters	Blinding of raters
Diagnostic criteria	Rater professions	Key conclusions of study authors
Inclusion criteria	Experience	Statistical tests
Exclusion criteria	Training	COSMIN criteria
Time between measurements	Whether patients were stable in the interim	

[a]COSMIN, Consensus-Based Standards for the Selection of Health Measurement Instruments.

"adequate," "doubtful," and "inadequate." Individual scores on the COSMIN "reliability" subitems were assessed and are included in Appendix 2 for completeness. COSMIN question 6.8, "other methodological flaws," was not assessed because of the subjectivity of the question. To minimize selection bias, studies were not excluded on the basis of methodological quality, as they were evaluated only in the reliability domain and the lowest score determined the overall quality in reliability.

We defined "reliability" as reproducibility of test values in repeated trials on the same individuals,[32] quantified by inter- and intrarater reliability. Consistency of outcomes recorded from 1 participant examined by the same observer multiple times was defined as intrarater reliability, while reproducibility of the score among observers was defined as interrater reliability.[4] Since the level of measurement of the Beighton score is not defined, researchers use different statistics to quantify these 2 values. Nominal and ordinal data were analyzed with the Cohen or weighted kappa (κ) coefficient,[50] which varies from –1 to 1. COSMIN criteria favor weighted kappa values, which penalize disagreements in terms of their seriousness, over unweighted kappa values, which treat disparities equally.[14,56]

Less rigorous expressions of inter- and intrarater reliability include percentage agreement and the Spearman rho. While percentage agreement is a direct measurement of the similarity between chosen values, it does not take into account the chance that scores were guessed[42] or the difference between more disparate scores. The Spearman rho expresses correlation between values on a scale of –1 to 1, with no known standards for reliability. This correlation reveals only how much values vary in relationship to each other, not the degree of agreement between them, allowing it to discount systematic differences.[9] These values were not considered adequate to express reliability according to COSMIN standards.

No transformation of reported values was required, except for simplifications detailed in the legend of Tables 1 and 2. No quantitative assessment of risk of bias across studies could be computed with the measures of reliability, and no additional quantitative analysis was performed.

TABLE 2
Strength of Agreement for the Kappa Coefficient and Intraclass Correlations[14,39,40,56]

Kappa Coefficient	Agreement	Intraclass Correlation	Reliability
$\leq$0	Poor	0.5	Poor
0.01-0.20	Slight	>0.5-0.75	Moderate
0.21-0.40	Fair	>0.75-0.9	Good
0.41-0.60	Moderate	>0.90	Excellent
0.61-0.80	Substantial		
0.81-1.00	Almost perfect		

RESULTS

The database search strategy yielded 1250 records. Three articles not identified by these searches were discovered by literature citation and added to the screen. After the screening process delineated in Figure 1, a total of 24 records were determined to meet inclusion criteria.[‡]

Table 3 includes characteristics of all included studies and their corresponding COSMIN criteria. All 24 studies selected for review were published in English and were observational studies with level 4 evidence. Of the 14 articles that explicitly express time intervals between measurements, the longest was 12 to 16 weeks,[5] with 12 of 14 reporting $\leq$2 weeks. A total of 1333 participants were examined for reliability of the Beighton score across included trials, with a reported mean ± SD age of 28.19 ± 17.34 years (range, 4-71 years). Of the 24 studies, 8 had populations <18 years old, and 14 included a higher proportion of women than men (640 female, 594 male, 273 unknown). Six studies included athletes in their participant population, and 8 comprised patients with pathological conditions. Seven studies used goniometers in their protocol.

Raters in at least 18 of the 24 studies were health care professionals (HCPs) or HCP students. Eight studies had physical therapists or physical therapy students as raters; 2 studies, orthopaedic surgeons; 3 studies, rheumatologists; and in the 5 other studies, other HCP disciplines that were not specified in the article. One study referred to its raters as "researchers." None of the studies included HCPs with equal years of experience. Half of the studies did not report the HCPs' years of experience at all. For the studies that did

[‡]References 1, 3, 5, 8, 11, 15, 18, 20–23, 27, 29, 31, 34, 35, 37, 46, 49, 57, 58, 62, 63, 65.

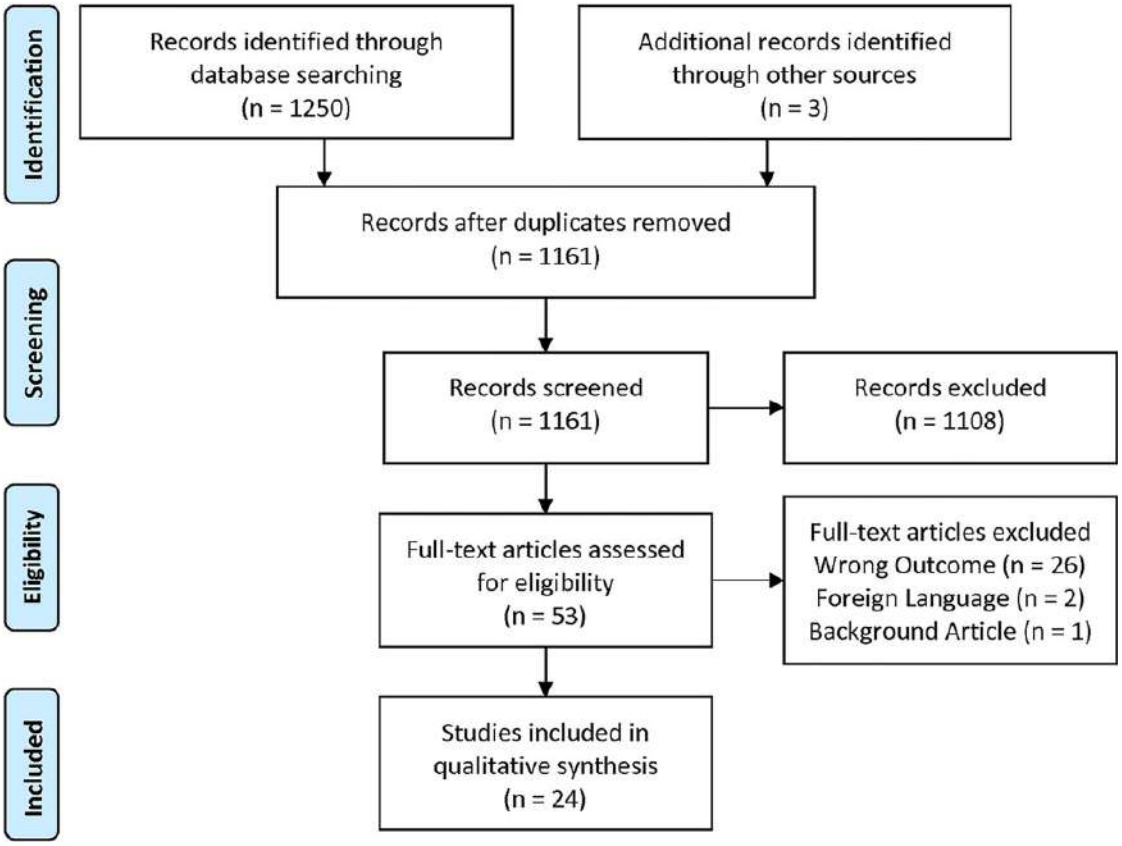

Figure 1. Flow diagram summarizing the literature search, screening, and review using the PRISMA (Preferred Reporting Items for Systematic Meta-Analyses) guidelines.

TABLE 3
Population Characteristics, Time Interval, Study Design, and Associated COSMIN Scores[a]

	Population Characteristics	Time Interval				Study Design					
Study (Year)	Sample, Age (y), Female Sex (%), DOP[b,c]	6.1[d]	Interrater	Intrarater	6.2[d]	Test Condition	No. of Raters	Rater Profession	Combined Rater Experience[c]	Rater Training	6.3[d]
Aartun (2014)[1]	111, 12-14, 46.8, middle school students	VG	<4 d	1-4 h	VG	5 item	2	Chiropractors	18 y	Standardization session	VG
Aslan (2006)[3]	72, 20.36 ± 1.24 (18-25), 40.20, undergraduate PT students	VG	<24 h	12.84 ± 7.41 d	VG	5 item + goniometer	2	PTs	21 y	2 h practice together	VG
Baumhauer (1995)[5]	21, 18-23, 57, intercollegiate athletes	VG	12-16 wk	NA	VG	5 item	2	NS	NS	NS	A
Boyle (2003)[8]	42, 25.4 ± 4.2 (15-45), 100, noninjured HS athletes and PT students	VG	15-60 min	6 ± 4 d	VG	5 item + goniometer	2	PTs	17 y	CME, trained with index	VG
Bulbena (1992)[11]	173, 43.98[e] NS, JHS with >5 Beighton system	VG	Consecutive	NA	D	5 item	2	Rheumatologists	Experienced	NS	A

(continued)

Table 3 (continued)

	Population Characteristics	Time Interval				Study Design					
Study (Year)	Sample, Age (y), Female Sex (%), DOP[b,c]	6.1[d]	Interrater	Intrarater	6.2[d]	Test Condition	No. of Raters	Rater Profession	Combined Rater Experience[c]	Rater Training	6.3[d]
Cooper (2018)[15]	50, 49 (22-60), 56, community members	VG	NS	1 wk	VG	5 item + goniometer	1	NS	NS	NS	A
Erdogan (2012)[18]	15, 31.8 (16-50), 59.15, treated for ingrown nails	VG	NS	NS	D	5 item + goniometer	2	Rheumatologists	NS	NS	A
Erkula (2005)[20]	50, 10.4 ± 1.2 (8-15)[f] 46.97, asymptomatic students	VG	2 wk	NA	VG	5 item	2	Orthopaedic surgeons	NS	NS	A
Evans (2012)[21]	30, 10.6 ± 2.3 (7-15), 65, asymptomatic podiatry clinic patients	VG	>2 h	>2 h	VG	5 item	2	Podiatrists	21 y	NS	A
Fritz (2005)[22]	38, 39.2 ± 11[f] 57,[f] history of lower back pain	VG	5 min	NA	VG	5 item	2	PTs	NS	NS	A
Glasoe (2002)[23]	30, 14-24, 100, athletes	VG	NS	NA	VG	5 item	2	NS	>6 y	NS	A
Hansen (2002)[27]	100, 9-13, NS, asymptomatic competitive athletes	VG	NS	NA	D	4/5, no fifth finger	4	2 rheumatologists 1 untrained physician	NS	Guided by illustrations	A
Hicks (2003)[29]	63, 36 (20-66), 60.30, patients with lower back pain	VG	>15 min	NA	VG	5 item	4	3 PT, 1 PT and chiropractor	20 y	Group review, 1 h practice	VG
Hirsch (2007)[31]	50, 38.3 ± 11.3 (20-60), 56, asymptomatic	VG	NS	24.6 d	VG	5 item + goniometer	2	Dentists	NS	Instructions, directed by orthopaedic surgeon	VG
Junge (2013)[34]	103, 7-8 and 10-12, 44[e] healthy school children	VG	<30 min	NA	VG		4	PT students	NS	Trained	VG
Juul-Kristensen (2007)[35]	40, 42.27 (18-71)[e] 68.33,[e] BJHS, EDS, back/ shoulder pain	VG	NS	NA	D	5 item	2	NS	NS	Trained per protocol	VG
Karim (2011)[37]	30, 24 (18-32), 100, contemporary professional dancers	VG	NS	NA	VG	5 item	4	1 PT, 3 PT students	30 y	PT trained students	VG
Naal (2014)[46]	55, 28.5 ± 4.1, 32.70, symptomatic FAI cases	VG	NS	NA	D	5 item	2	Clinicians	NS	NS	A
Pitetti (2015)[49]	25, 13.3 ± 2.9, 44, intellectually disabled	VG	3-4 wk	NA	VG	5 item + goniometer	2	DPT students	None	Peer supportive learning	VG
Smith (2012)[57]	5, 27, 100, patellar instability patients	VG	<1 d	30 min	VG	5 item	5	Orthopaedic surgeons	125 y	Familiarized	VG
Tarara (2014)[58]	19, 20.3 ± 1.2 (male), 19.8 ± 1.0 (female), 57.89, club athletes	VG	<2.5 h	4-7 d	VG	5 item	3	1 clinician and 2 novice students	22 y	Prior reading, 1 h training and questions	VG

(continued)

Table 3 (continued)

Study (Year)	Population Characteristics: Sample, Age (y), Female Sex (%), DOP[b,c]	Time Interval: 6.1[d]	Time Interval: Interrater	Time Interval: Intrarater	Time Interval: 6.2[d]	Study Design: Test Condition	Study Design: No. of Raters	Study Design: Rater Profession	Study Design: Combined Rater Experience[c]	Study Design: Rater Training	Study Design: 6.3[d]
Vaishya (2013)[62]	300, 24.6 ± 0.9, 36.67, postoperative ACL reconstruction and controls	VG	NS	NA	D	5 item	2	NS	NS	NS	A
Vallis (2015)[63]	36, 22.7 (18-32), 75, asymptomatic PT and OT students	VG	<1 d, 1 wk	NA	VG	5 item + goniometer	2	Researchers	NS	Teaching session	VG
van der Giessen (2001)[65]	48, 4-12, 48.9[f] primary schoolchildren	VG	NS	NA	D	5 item	2	PT students	1 mo	Professional PT trained students	VG

[a]ACL, anterior cruciate ligament; BJHS, benign joint hypermobility syndrome; CME, continuing medical education; DOP, description of participants; DPT, doctorate of physical therapy; EDS, Ehlers-Danlos syndrome; FAI, femoroacetabular impingement; HS, high school; JHS, joint hypermobility syndrome; NA, not available/applicable; NS, not specified; OT, occupational therapy; PT, physical therapy.

[b]Age reported as mean ± SD or range.

[c]Calculated.

[d]COSMIN criterion (Consensus-Based Standards for the Selection of Health Measurement Instruments; see Appendix 2 for details). Scoring: VG = very good, A = adequate, D = doubtful, I = inadequate.

[e]Weighted average of groups or 2-phase studies.

[f]Demographics of larger sample, of which reliability population is a subgroup.

report years of experience, the numbers for each HCP were summed to reach combined total years for Table 3.

Table 4 includes measures of reliability in each study and the corresponding COSMIN criteria. Because the study designs, participants, interventions, and reported outcome measures varied markedly, results were synthesized in a qualitative manner, and pooled means could not be determined. Because 3 studies included reliability statistics for >1 cutoff score (ie, ≥4/9 and composite), the included 24 articles reported interrater reliability values for 27 cutoff scores. For interrater reliability, 5 of the 27 scoring cutoffs were ≥4 of 9; 3 were ≥5 of 9; 13 were composite (total of 9 points); 4 used each item in the Beighton score; and 1 used a modified composite scale. Intrarater reliability was expressed for 10 total cutoff values: 3 were ≥4 of 9; 1 was ≥5 of 9; 5 included composite values; and 1 calculated a score for each item. Of the 8 studies that utilized intraclass correlation (ICC) to express interrater reliability, 1 found an excellent value; 4, good; 1, moderate to good; and 2, moderate. Of the 19 kappa values or ranges for interrater reliability, 3 were almost perfect; 6 were substantial; 2 were moderate; 1 was poor; and the others ranged between scales. Of the 7 ranges, 3 crossed between substantial and almost perfect, while the other 4 varied among lower ratings. Three studies used percentage agreement values, and 3 studies used the Spearman rho to demonstrate interrater reliability. For interrater reliability, 5 of 8 (62.5%) ICCs and 12 of 19 (63.2%) kappa values were better than moderate. Of the 13 intrarater values provided, 3 were ICC; 7 were kappa; 2 were percentage agreement; and 1 was a Spearman rho. All 3 ICC values for intrarater reliability were excellent. For the 7 kappa values and ranges, 2 were almost perfect; 2, substantial; 1, fair; and 2 had scores varying from substantial to almost perfect.

Out of the 168 COSMIN questions in the reliability domain across all studies, 79 (47%) were "very good"; 29 (17%), "adequate"; 24 (14%), "doubtful"; 1, "inadequate"; and 35 (21%) did not apply. Utilizing the COSMIN "worse score counts" principle, we determined that 1 (4%) study met "very good" criteria[29]; 7 (29%) met "adequate"[3,21,22,31,57,58,63]; 15 (63%) met "doubtful"[§]; and 1 (4%) met "inadequate"[5] for overall risk of bias in the reliability domain. Eight (33.33%) studies utilized ICC, and 16 (66.66%) comprised 19 kappa statistics to express interrater reliability, of which 4 (25%) used weighted kappa values. Of the 12 articles that included unweighted kappa values, 6 received an overall score of "doubtful," which was attributed only to question 6.6, regarding use of weighted kappa,[44] when they otherwise would have received "adequate" or "very good" overall. Of the 24 included studies, 7 did not report an explicit time interval between reliability measurements. However, 6 of the 7 had another doubtful measure, which means that question 6.2, regarding the appropriateness of the time interval,[44] did not greatly affect the overall score for most studies.

DISCUSSION

This systematic review has demonstrated high inter- and intrarater reliability for the Beighton score in individuals

[§]References 1, 8, 11, 15, 18, 20, 23, 27, 34, 35, 37, 46, 49, 62, 65.

TABLE 4
Inter- and Intrarater Reliability and Associated COSMIN Scores[a]

Study (Year)	Cutoff Score	Reliability, Mean (95% CI)		COSMIN Item			
		Interrater	Intrarater	6.4	6.5	6.6	6.7
Aartun (2014)[1]	≥4/9	κ = 0.65 (0.33 to 0.97)	κ = 0.66-1 (0.03 to 1)	NA	VG	D	A
	≥5/9	κ = 0.56 (0.11 to 1.00)	κ = 1				
Aslan (2006)[3]	Composite	ICC = 0.82 Agreement = 42%	ICC = 0.92 Agreement = 43%	A	NA	NA	NA
Baumhauer (1995)[5]	Composite	ρ = 1		NA	I	D	A
Boyle (2003)[8]	Composite	ρ = 0.87 Agreement = 51%	ρ = 0.86 Agreement = 69%	D	NA	NA	NA
Bulbena (1992)[11]	Each item	κ = 0.79-0.93		D	VG	D	NA
Cooper (2018)[15]	≥4/9	κ = 0.96[b] (0.87 to 1.00)	κ = 1	NA	VG	D	A
Erdogan (2012)[18]	Each item	κ = 0.71-1.0	κ = 0.81-1.0	NA	VG	D	A
Erkula (2005)[20]		ρ = 0.86	ρ = 0.62	D	NA	NA	NA
Evans (2012)[21]	Composite	ICC = 0.73	ICC = 0.96-0.98	VG	NA	NA	NA
Fritz (2005)[22]	Composite	ICC = 0.72 (0.50 to 0.85)		VG	NA	NA	NA
Glasoe (2002)[23]	Composite	κ = 0.7		NA	VG	D	A
Hansen (2002)[27]	≥4/9	κ = 0.44-0.82		D	VG	D	A
Hicks (2003)[29]	Composite	ICC = 0.79 (0.68 to 0.87)		VG	NA	NA	NA
Hirsch (2007)[31]	≥4/9	ICC >0.84	ICC > 0.89	A	NA	NA	NA
Junge (2013)[34]	Each item [c]	κ = 0.49-0.94, 0.30-0.84		NA	VG	D	A
	≥5/9[c]	κ = 0.64, 0.59[d]					
Juul-Kristensen (2007)[35]	Composite	ICC = 0.91		VG	VG	D	A
	≥5/9	κ = 0.66 (0.30 to 1.02) 0.74 (0.46 to 1.02)[d]					
Karim (2011)[37]	NS	κ = 0.6 Agreement = 54%-100%		NA	VG	D	NA
Naal (2014)[46]	Composite	κ = 0.82[b] (0.72 to 0.91)		NA	VG	VG	VG
Pitetti (2015)[49]	Composite	ICC = 0.88		A	VG	D	A
	Each item	κ = 0.45-0.80					
Smith (2012)[57]	Composite	κ = 0.00 (−0.16 to 0.17)	κ = 0.25 (0.03 to 0.51)	NA	VG	VG	A
Tarara (2014)[57]	Modified composite[e]	κ = 0.64-0.69[f] κ = 0.72[g] (0.62 to 0.82)	Expert: κ = 0.69 (0.46 to 0.92) Novice: κ = 0.72-0.73 ([0.53-0.90] to [0.58-0.89])	NA	VG	VG	A
Vaishya (2013)[62]	≥4/9	κ = 0.7		NA	VG	D	A
Vallis (2015)[63]	Composite	ICC = 0.72-0.80 ([0.51-0.84] to [0.64-0.89]) κ = 0.71-0.82 ([0.67-0.90] to [0.50-0.84])		A	VG	VG	A
van der Giessen (2001)[65]	Composite	κ = 0.81		NA	VG	D	A

[a]A, adequate; COSMIN, Consensus-Based Standards for the Selection of Health Measurement Instruments; D, doubtful; I, inadequate; ICC, intraclass correlation; NA, not available/applicable; VG, very good.
[b]Observer-participant reliability.
[c]Percentage agreement omitted.
[d]For 2 distinct methods of performing Beighton score.
[e]Modified composite scale: 0 = pain with test, 1 = 8-9 points, 2 = 6-7 points, 3 = 4-5 points, 4 = 2-3 points, 5 = 0-1 points.
[f]Expert-novice rater reliability.
[g]Novice-novice rater reliability.

with and without hypermobility in a variety of clinical conditions. As demonstrated by the data derived from Table 3, varying time conditions, population characteristics, measurement tools, measurer education and training, and the Beighton score cutoff did not greatly influence the reliability of this test. Most studies demonstrated substantial to almost perfect interrater reliability values. Intrarater reliability was excellent or almost perfect in more than half of analyzed investigations. The quality of analyzed evidence was adequate, in contrast to findings in previous systematic reviews.[35]

The increased mobility seen in patients with an elevated Beighton score is of importance for the clinician. Generalized joint hypermobility is a risk factor for many musculoskeletal conditions, such as multidirectional shoulder instability,[54] hip instability,[12] femoroacetabular impingement,[46,64] hip dysplasia,[2,7] ACL injury,[60,62] flatfoot,[45] ankle sprains,[16] and many others. Clinicians should have a high

index of suspicion for these conditions in this population. Knowledge of hypermobility influences patient selection for surgical versus nonsurgical treatments, the actual surgical technique employed, and the expected prognosis and outcome with respect to risks for recurrence of symptoms (which may vary along a spectrum of instability).[19] This is important in the clinical setting for practitioners to avoid unnecessary imaging or interventions or the misdiagnosis of chronic pain.[66]

Patients with hypermobility may warrant more aggressive rehabilitation or injury prevention protocols. Owing to the higher incidence of joint instability in patients with hypermobility, it has been suggested that these patients undergo prolonged strengthening, proprioception, and generalized conditioning programs when considering initial nonoperative treatment.[66] Additionally, considerations in operative intervention may change with the knowledge of a patient's hypermobility status. For instance, a surgeon might consider an open inferior capsular shift versus arthroscopic capsular plication for the hypermobile shoulder, or a surgeon may consider using a patellar tendon autograft over hamstring tendon autograft in ACL reconstruction[38] to ensure greater stability postoperatively. Arthroscopic hip preservation surgeons may employ greater degrees of capsular plication and/or inferior capsular shift in patients undergoing FAI syndrome and labral injury surgical treatment.[61] Even trauma and arthroplasty surgeons should consider a patient's hypermobility status. Patients with hypermobility have been found to have lower bone density[25,47] than controls, which leaves them at greater risk for fixation and implant failure and fracture. Postoperative protocols may need to be adjusted for this population to address the increased laxity. Thus, use of a reliable system, such as the Beighton score, for identifying these patients is essential to providing the most comprehensive musculoskeletal care.

Limitations of the present study include the quality of studies available in the literature, the failure of studies to include time intervals between intrarater measures, reporting bias, and lack of rater standardization or comparison. Studies that did not include time intervals between intrarater measures resulted in a summary COSMIN score of "doubtful." Laxity may change in an individual over a period of decades[3,11,16,66]; however, it does not change over short periods. Thus, the omission of time intervals should not negatively affect a clinician's evaluation of the evidence supporting inter- and intrarater reliability of the Beighton score. Additionally, score reporting is subject to publication bias and selective reporting because reliability may be reported by composite score, individual measurement score, or cutoff score. This may influence authors to choose the reporting measure with the most desirable outcomes. Studies that measure interrater reliability risk underestimating it when raters are not properly standardized. Using raters with unequal experience may result in artificially low interrater statistics. All studies in the present review used raters with different levels of experience; thus, it is likely that under standardized conditions the interrater reliability may be higher. No one study utilized raters of different professions; therefore, the discrepancy in Beighton score reliability among health care disciplines cannot be evaluated by this study.

CONCLUSION

The Beighton score is a reliable clinical assessment tool that shows acceptable reliability when used by raters of any background or experience level. Studies demonstrate immense variability in participant population, study design, time interval, and rater experience yet consistently report substantial to excellent inter- and intrarater reliability. While individual components of risk of bias among studies also demonstrated large discrepancy, most of the items were adequate to very good.

REFERENCES

1. Aartun E, Degerfalk A, Kentsdotter L, Hestbaek L. Screening of the spine in adolescents: inter- and intra-rater reliability and measurement error of commonly used clinical tests. *BMC Musculoskelet Disord*. 2014;15(1):37.
2. Adib N, Davies K, Grahame R, Woo P, Murray KJ. Joint hypermobility syndrome in childhood: a not so benign multisystem disorder? *Rheumatology*. 2005;44(6):744-750.
3. Aslan UB, Çelik E, Cavlak U, Akdağ B. Evaluation of interrater and intrarater reliability of Beighton and Horan Joint Mobility Index. *Fiz Rehabil*. 2006;17(3):113-119.
4. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med*. 1998;26(4):217-238.
5. Baumhauer JF, Alosa DM, Renström PAFH, Trevino S, Beynnon B. Test-retest reliability of ankle injury risk factors. *Am J Sports Med*. 1995;23(5):571-574.
6. Beighton P, Solomon L, Soskolne CL. Articular mobility in an African population. *Ann Rheum Dis*. 1973;32(5):413-418.
7. Bilsel K, Ceylan HH, Yıldız F, Erden T, Toprak A, Tuncay I. Acetabular dysplasia may be related to global joint hyperlaxity. *Int Orthop*. 2016;40(5):885-889.
8. Boyle KL, Witt P, Riegger-Krugh C. Intrarater and interrater reliability of the Beighton and Horan Joint Mobility Index. *J Athl Train*. 2003;38(4):281-285.
9. Bruton A, Conway JH, Holgate ST. Reliability: what is it, and how is it measured? *Physiotherapy*. 2000;86(2):94-99.
10. Buchanan G, Torres L, Czarkowski B, Giangarra CE. Current concepts in the treatment of gross patellofemoral instability. *Int J Sports Phys Ther*. 2016;11(6):867.
11. Bulbena A, Duró JC, Porta M, Faus S, Vallescar R, Martín-Santos R. Clinical assessment of hypermobility of joints: assembling criteria. *J Rheumatol*. 1992;19(1):115.
12. Canham CD, Domb BG, Giordano BD. Atraumatic hip instability. *JBJS Rev*. 2016;4(5):01874474-201605000-00001.
13. Castori M, Tinkle B, Levy H, Grahame R, Malfait F, Hakim A. A framework for the classification of joint hypermobility and related conditions. *Am J Med Genet Part C Semin Med Genet*. 2017;175(1):148-157.
14. Cohen J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol Bull*. 1968;70(4):213-220.
15. Cooper DJ, Scammell BE, Batt ME, Palmer D. Development and validation of self-reported line drawings of the modified Beighton score for the assessment of generalised joint hypermobility. *BMC Med Res Methodol*. 2018;18(1):11.
16. Decoster LC, Bernier JN, Lindsay RH, Vailas JC. Generalized joint hypermobility and its relationship to injury patterns among NCAA lacrosse players. *J Athl Train*. 1999;34(2):99.

17. Denteneer L, Stassijns G, De Hertogh W, Truijen S, Van Daele U. Inter- and intrarater reliability of clinical tests associated with functional lumbar segmental instability and motor control impairment in patients with low back pain: a systematic review. *Arch Phys Med Rehabil*. 2017;98(1):151-164.e6.
18. Erdogan FG, Tufan A, Guven M, Goker B, Gurler A. Association of hypermobility and ingrown nails. *Clin Rheumatol*. 2012;31(9): 1319-1322.
19. Ericson WB, Wolman R. Orthopaedic management of the Ehlers-Danlos syndromes. *Am J Med Genet Part C Semin Med Genet*. 2017;175(1):188-194.
20. Erkula G, Kiter AE, Kilic BA, Er E, Demirkan F, Sponseller PD. The relation of joint laxity and trunk rotation. *J Pediatr Orthop B*. 2005; 14(1):38-41.
21. Evans AM, Rome K, Peet L. The foot posture index, ankle lunge test, Beighton scale and the lower limb assessment score in healthy children: a reliability study. *J Foot Ankle Res*. 2012;5(1):1.
22. Fritz JM, Piva SR, Childs JD. Accuracy of the clinical examination to predict radiographic instability of the lumbar spine. *Eur Spine J*. 2005; 14(8):743-750.
23. Glasoe WM, Allen MK, Kepros T, Stonewall L, Ludewig PM. Dorsal first ray mobility in women athletes with a history of stress fracture of the second or third metatarsal. *J Orthop Sports Phys Ther*. 2002; 32(11):560-565.
24. Grahame R. The need to take a fresh look at criteria for hypermobility. *J Rheumatol*. 2007;34(4):664.
25. Gulbahar S, Şahin E, Baydar M, et al. Hypermobility syndrome increases the risk for low bone mass. *Clin Rheumatol*. 2006;25(4): 511-514.
26. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quant Methods Psychol*. 2012;8(1): 23-34.
27. Hansen A, Damsgaard R, Kristensen JH, Bagger J, Remvig L. Inter-examiner reliability of selected tests for hypermobility. *J Orthop Med*. 2002;24(2):48-51.
28. Harris JD, Delgado DA, Dong D, Bockhorn L. Beighton score inter-rater reliability: a systematic review. Published 2018. http://www.crd.york.ac.uk/PROSPERO/display_record.php?ID=CRD42018081703
29. Hicks GE, Fritz JM, Delitto A, Mishock J. Interrater reliability of clinical examination measures for identification of lumbar segmental instability. *Arch Phys Med Rehabil*. 2003;84(12):1858-1864.
30. Higgins J. Cochrane handbook for systematic reviews of interventions. Published 2011. http://www.cochrane-handbook.org/
31. Hirsch C, Hirsch M, John MT, Bock JJ. Reliabilität der Beighton-Skala zur Bestimmung der allgemeinen Gelenküberbeweglichkeit durch zahnärztliche Untersucher. *J Orofac Orthop*. 2007;68(5):342-352.
32. Hopkins WG. Measures of reliability in sports medicine and science. *Sports Med*. 2000;30(1):1-15.
33. Hopkins WG, Marshall SW, Batterham AM, Hanin J. Progressive statistics for studies in sports medicine and exercise science. *Med Sci Sports Exerc*. 2009;41(1):3-12.
34. Junge T, Jespersen E, Wedderkopp N, Juul-Kristensen B. Inter-tester reproducibility and inter-method agreement of two variations of the Beighton test for determining generalised joint hypermobility in primary school children. *BMC Pediatr*. 2013;13(1):214.
35. Juul-Kristensen B, Røgind H, Jensen DV, Remvig L. Inter-examiner reproducibility of tests and criteria for generalized joint hypermobility and benign joint hypermobility syndrome. *Rheumatology*. 2007; 46(12):1835-1841.
36. Juul-Kristensen B, Schmedling K, Rombaut L, Lund H, Engelbert RHH. Measurement properties of clinical assessment methods for classifying generalized joint hypermobility—a systematic review. *Am J Med Genet Part C Semin Med Genet*. 2017;175(1):116-147.
37. Karim A, Millet V, Massie K, Olson S, Morganthaler A. Inter-rater reliability of a musculoskeletal screen as administered to female professional contemporary dancers. *Work*. 2011;40(3):281.
38. Kim SJ, Kumar P, Kim SH. Anterior cruciate ligament reconstruction in patients with generalized joint laxity. *Clin Orthop Surg*. 2010;2(3): 130-139.
39. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2): 155-163.
40. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159.
41. Larson CM, Bedi A, Dietrich ME, et al. Generalized hypermobility, knee hyperextension, and outcomes after anterior cruciate ligament reconstruction: prospective, case-control study with mean 6 years follow-up. *Arthroscopy*. 2017;33(10):1852-1858.
42. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22(3):276-282.
43. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA statement. *BMJ*. 2009;339(7716):B2535.
44. Mokkink LB, de Vet HCW, Prinsen CAC, et al. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Qual Life Res*. 2018;27(5):1171-1179.
45. Murray KJ. Hypermobility disorders in children and adolescents. *Best Pract Res Clin Rheumatol*. 2006;20(2):329-351.
46. Naal FD, Hatzung G, Müller A, Impellizzeri F, Leunig M. Validation of a self-reported Beighton score to assess hypermobility in patients with femoroacetabular impingement. *Int Orthop*. 2014;38(11):2245-2250.
47. Nijs J, Van Essche E, De Munck M, Dequeker J. Ultrasonographic, axial, and peripheral measurements in female patients with benign hypermobility syndrome. *Calcif Tissue Int*. 2000;67(1):37-40.
48. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev*. 2016;5(1):210.
49. Pitetti K, Miller RA, Beets MW. Measuring joint hypermobility using the Beighton scale in children with intellectual disability. *Pediatr Phys Ther*. 2015;27(2):143-150.
50. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. *Clin Rehabil*. 1998; 12(3):187-199.
51. Remvig L, Flycht L, Christensen KB, Juul-Kristensen B. Lack of consensus on tests and criteria for generalized joint hypermobility, Ehlers-Danlos syndrome: hypermobile type and joint hypermobility syndrome. *Am J Med Genet A*. 2014;164(3):591-596.
52. Remvig L, Jensen DV, Ward RC. Epidemiology of general joint hypermobility and basis for the proposed criteria for benign joint hypermobility syndrome: review of the literature. *J Rheumatol*. 2007;34(4): 804-809.
53. Ropars M, Cretual A, Thomazeau H, Kaila R, Bonan I. Volumetric definition of shoulder range of motion and its correlation with clinical signs of shoulder hyperlaxity: a motion capture study. *J Shoulder Elbow Surg*. 2015;24(2):310-316.
54. Saccomanno MF, Fodale M, Capasso L, Cazzato G, Milano G. Reconstruction of the coracoclavicular and acromioclavicular ligaments with semitendinosus tendon graft: a pilot study. *Joints*. 2014; 2(1):6-14.
55. Sacks HA, Prabhakar P, Wessel LE, et al. Generalized joint laxity in orthopaedic patients: clinical manifestations, radiographic correlates, and management. *J Bone Joint Surg Am*. 2019;101(6):558-566.
56. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther*. 2005;85(3): 257-268.
57. Smith TO, Clark A, Neda S, et al. The intra- and inter-observer reliability of the physical examination methods used to assess patients with patellofemoral joint instability. *Knee*. 2012;19(4):404-410.
58. Tarara DT, Hegedus EJ, Taylor JB. Real-time test-retest and interrater reliability of select physical performance measures in physically active college-aged students. *Int J Sports Phys Ther*. 2014;9(7):874.
59. Tinkle B, Castori M, Berglund B, et al. Hypermobile Ehlers-Danlos syndrome (aka Ehlers-Danlos syndrome type III and Ehlers-Danlos syndrome hypermobility type): clinical description and natural history. *Am J Med Genet Part C Semin Med Genet*. 2017;175(1):48-69.
60. Uhorchak JM, Scoville CR, Williams GN, Arciero RA, St Pierre P, Taylor DC. Risk factors associated with noncontact injury of the anterior cruciate ligament: a prospective four-year evaluation of 859 West Point cadets. *Am J Sports Med*. 2003;31(6):831-842.

61. Ukwuani GC, Waterman BR, Nwachukwu BU, et al. Return to dance and predictors of outcome after hip arthroscopy for femoroacetabular impingement syndrome. *Arthroscopy*. 2019;35(4):1101-1108.
62. Vaishya R, Hasija R. Joint hypermobility and anterior cruciate ligament injury. *J Orthop Surg (Hong Kong)*. 2013;21(2):182-184.
63. Vallis A, Wray A, Smith T. Inter- and intra-rater reliabilities of the Beighton score compared to the contompasis score to assess generalised joint hypermobility. *Myopain*. 2015;23(1-2):21-27.
64. Weber AE, Bedi A, Tibor LM, Zaltz I, Larson CM. The hyperflexible hip: managing hip pain in the dancer and gymnast. *Sports Health*. 2015;7(4):346-358.
65. van der Giessen LJ, Liekens D, Rutgers KJ, Hartman A, Mulder PG, Oranje AP. Validation of Beighton score and prevalence of connective tissue signs in 773 Dutch children. *J Rheumatol*. 2001;28(12):2726.
66. Wolf JM, Cameron KL, Owens BD. Impact of joint laxity and hypermobility on the musculoskeletal system. *J Am Acad Orthop Surg*. 2011;19(8):463-471.

APPENDIX 1

SEARCH STRATEGIES

1. Medline, Ovid

1. beighton.ti,ab.
2. exp Joint Instability/
3. ((joint adj1 (laxity or instability)) or hypermobil*).ti,ab.
4. or/1-4
5. exp "Reproducibility of Results"/
6. exp Observer Variation/
7. (Reproducibil* or reliabil*).ti,ab.
8. (observer adj1 variation).ti,ab.
9. (interrater or intrarater or ((intra* or inter*) adj1 (test* or observ* or reliabil* or rater*))).ti,ab.
10. ((disagreement or agreement) adj4 score*).ti,ab.
11. or/5-10
12. 4 and 11
13. animals/ not humans/

2. Embase, Ovid

1. beighton.ti,ab.
2. exp joint laxity/ or exp joint hypermobility/ or exp joint instability/
3. ((joint adj1 (laxity or instability)) or hypermobil*).ti,ab.
4. or/1-3
5. reproducibility/
6. exp observer variation/
7. (Reproducibil* or reliabil*).ti,ab.
8. (observer adj1 variation).ti,ab.
9. (interrater or intrarater or ((intra* or inter*) adj1 (test* or observ* or reliabil* or rater*))).ti,ab.
10. ((disagreement or agreement) adj4 score*).ti,ab.
11. or/5-10
12. 4 and 11
13. animal/ not human/
14. 12 not 13

3. CINAHL

- TI beighton OR AB beighton OR (MH "Joint Instability+") OR TI (((joint n1 (laxity or instability)) or hypermobil*)) OR AB (((joint n1 (laxity or instability)) or hypermobil*))
- AND
- (MH "Reproducibility of Results") OR ((MH "Interrater Reliability") OR (MH "Reliability") OR (MH "Reliability and Validity") OR (MH "Intrarater Reliability")) OR TI (((disagreement or agreement) n4 score*) or (interrater or intrarater or ((intra* or inter*) n1 (test* or observ* or reliabil* or rater*))) or (Reproducibil* or reliabil*) or (observer n1 variation)) OR AB (((disagreement or agreement) n4 score*) or (interrater or intrarater or ((intra* or inter*) n1 (test* or observ* or reliabil* or rater*))) or (Reproducibil* or reliabil*) or (observer n1 variation))
- English Language; Peer Reviewed; Exclude MEDLINE records

4. SPORTDiscus

- DE "INTER-observer reliability" OR TI (((disagreement or agreement) n4 score*) or (interrater or intrarater or ((intra* or inter*) n1 (test* or observ* or reliabil* or rater*))) or (Reproducibil* or reliabil*) or (observer n1 variation)) OR AB (((disagreement or agreement) n4 score*) or (interrater or intrarater or ((intra* or inter*) n1 (test* or observ* or reliabil* or rater*))) or (Reproducibil* or reliabil*) or (observer n1 variation))
- AND
- DE "HYPERMOBILITY of joints" OR (TI beighton OR AB beighton OR TI (((joint n1 (laxity or instability)) or hypermobil*)) OR AB (((joint n1 (laxity or instability)) or hypermobil*))

APPENDIX 2

COSMIN RISK-OF-BIAS CHECKLIST: RELIABILITY SECTION

BOX 6
Reliability

	Very Good	Adequate	Doubtful	Inadequate	Not applicable
Design requirements					
1 Were patients stable in the interim period on the construct to be measured?	Evidence provided that patients were stable	Assumable that patients were stable	Unclear if patients were stable	Patients were NOT stable	
2 Was the time interval appropriate?	Time interval appropriate		Doubtful whether time interval was appropriate or time interval was not stated	Time interval NOT appropriate	
3 Were the test conditions similar for the measurements (eg type of administration, environment, instructions)?	Test conditions were similar (evidence provided)	Assumable that test conditions were similar	Unclear if test conditions were similar	Test conditions were NOT similar	
Statistical methods					
4 For continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	ICC calculated and model or formula of the ICC is described	ICC calculated but model or formula of the ICC not described or not optimal. Pearson or Spearman correlation coefficient calculated with evidence provided that no systematic change has occurred	Pearson or Spearman correlation coefficient calculated WITHOUT evidence provided that no systematic change has occurred or WITH evidence that systematic change has occurred	No ICC or Pearson or Spearman correlations calculated	Not applicable
5 For dichotomous/ nominal/ordinal scores: Was kappa calculated?	Kappa calculated			No kappa calculated	Not applicable
6 For ordinal scores: Was a weighted kappa calculated?	Weighted Kappa calculated		Unweighted Kappa calculated or not described		Not applicable
7 For ordinal scores: Was the weighting scheme described? eg linear, quadratic	Weighting scheme described	Weighting scheme NOT described			Not applicable
Other					
8 Were there any other important flaws in the design or statistical methods of the study?	No other important methodological flaws		Other minor methodological flaws	Other important methodological flaws	

From Mokkink LB, de Vet HCW, Prinsen CAC, et al. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Qual Life Res.* 2018;27(5):1171-1179.[44]