RESEARCH REVIEW

Measurement Properties of Clinical Assessment Methods for Classifying Generalized Joint Hypermobility—A Systematic Review

BIRGIT JUUL-KRISTENSEN,* KAROLINE SCHMEDLING, LIES ROMBAUT, HANS LUND, AND RAOUL H. H. ENGELBERT

The purpose was to perform a systematic review of clinical assessment methods for classifying Generalized Joint Hypermobility (GJH), evaluate their clinimetric properties, and perform the best evidence synthesis of these methods. Four test assessment methods (Beighton Score [BS], Carter and Wilkinson, Hospital del Mar, Rotes-Querol) and two questionnaire assessment methods (Five-part questionnaire [5PQ], Beighton Score-self reported [BS-self]) were identified on children or adults. Using the Consensus-based Standards for selection of health Measurement Instrument (COSMIN) checklist for evaluating the methodological quality of the identified studies, all included studies were rated "fair" or "poor." Most studies were using BS, and for BS the reliability most of the studies showed limited positive to conflicting evidence, with some shortcomings on studies for the validity. The three other test assessment methods lack satisfactory information on both reliability and validity. For the questionnaire assessment methods, 5PQ was the most frequently used, and reliability showed conflicting evidence, while the validity had limited positive to conflicting evidence compared with test assessment methods. For BS-self, the validity showed unknown evidence compared with test assessment methods. In conclusion, following recommended uniformity of testing procedures, the recommendation for clinical use in adults is BS with cut-point of 5 of 9 including historical information, while in children it is BS with cut-point of at least 6 of 9. However, more studies are needed to conclude on the validity properties of these assessment methods, and before evidence-based recommendations can be made for clinical use on the "best" assessment method for classifying GJH. © 2017 Wiley Periodicals, Inc.

KEY WORDS: Beighton tests; Carter and Wilkinson; Rotes-Querol; Hospital del Mar; five-part questionnaire; self-reported; clinimetrics; quality assessment; COSMIN; best evidence synthesis

How to cite this article: Juul-Kristensen B, Schmedling K, Rombaut L, Lund H, Engelbert RHH. 2017. Measurement properties of clinical assessment methods for classifying generalized joint hypermobility—A systematic review. Am J Med Genet Part C Semin Med Genet 175C:116–147.

INTRODUCTION

Generalized joint hypermobility (GJH) is relatively common, occurring in about 2–57% of different populations [Remvig et al., 2007b]. Important reasons for this may be the use of many different clinical assessment methods and criteria for classification and interpretation of GJH by these clinical assessment methods [Remvig et al., 2007a,b]. GJH is characterized by an ability to exceed the joints beyond the normal range of motion in multiple joints, either congenital or acquired [Remvig et al., 2011]. Many individuals with GJH are asymptomatic, which also makes it difficult to accurately estimate

Dr. Birgit Juul-Kristensen, Associate professor, PT, Department of Sports Sciences and Clinical Biomechanics, Research Unit of Musculoskeletal Function and Physiotherapy, University of Southern Denmark, Odense, Denmark.

Karoline Schmedling, M.Sc., PT, Department of Health Sciences, Institute of Occupational Therapy, Physiotherapy and Radiography, Bergen University College, Bergen, Norway.

Dr. Lies Rombaut, Postdoctoral researcher, PT, Center for Medical Genetics, Ghent University Hospital, Ghent, Belgium.

Dr. Hans Lund, Associate professor, PT, Department of Sports Sciences and Clinical Biomechanics, Research Unit of Musculoskeletal Function and Physiotherapy, University of Southern Denmark, Odense, Denmark; SEARCH (Synthesis of Evidence and Research), Department of Sports Science and Clinical Biomechanics, University of Southern Denmark, Odense, Denmark; Center for Evidence-Based Practice, Bergen University College, Bergen, Norway.

Dr. Raoul Engelbert, Department of Rehabilitation, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands; ACHIEVE, Faculty of Health, Center for Applied Research, University of Applied Sciences, Amsterdam, the Netherlands.

Conflicts of interest: There are no other financial interests that any of the authors may have, which could create a potential conflict of interest or the appearance of a conflict of interest with regards to the work.

*Correspondence to: Birgit Juul-Kristensen, Research Unit for Musculoskeletal Function and Physiotherapy, Institute of Sports Science and Clinical Biomechanics, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark. E-mail: bjuul-kristensen@health.sdu.dk

DOI 10.1002/ajmg.c.31540

Article first published online in Wiley Online Library (wileyonlinelibrary.com).

© 2017 Wiley Periodicals, Inc.

the number of people with this condition, as they are not recorded in the health system.

When GJH is accompanied with symptoms, it is defined as a health-related disorder, for example, Joint Hypermobility Syndrome (JHS) or the Ehlers–Danlos Syndrome—Hypermobile Type (hEDS) with several complications as described below. The two conditions (JHS and hEDS) have very close overlap to the point of being clinically indistinguishable [Tinkle et al., 2009; Remvig et al., 2011], and in the present study it is referred to as JHS/hEDS. The condition of JHS/hEDS can be defined as an under- and often misdiagnosed heritable connective tissue disorder, characterized generally by GJH, complications of joint instability, musculoskeletal pain, skin involvement, and reduced quality of life [Rombaut et al., 2010; Castori et al., 2014; Scheper et al., 2016]. Until now, JHS is diagnosed by the Brighton tests and criteria [Grahame et al., 2000], and hEDS by the Villefranche criteria [Beighton et al., 1998], both including the Beighton scoring (BS) system of nine tests for assessment of GJH [Beighton et al., 1973].

BS consists of four bilateral tests and one test including low back and lower extremities (first finger opposition, fifth finger extension, elbow extension, knee extension, and back forward bending), with scores ranging from 0 to 9. Influencing factors on BS are age, gender, ethnicity, and physical fitness [Remvig et al., 2007b; Tinkle et al., 2009]. For adults, a cut-point of 4/9 for GJH is included in the Brighton criteria for JHS [Grahame et al., 2000], while 5/9 for GJH is the criteria for hEDS in the Villefranche criteria [Beighton et al., 1998]. For children, there is no consensus on a specific cut-point for GJH, but cut-points of 5/9, 6/9, and 7/9 have been suggested [Jansson et al., 2004]. A previous review has listed different test assessment methods of which the Beighton score [Beighton et al., 1973] was most frequently used. The review concluded that reproducibility of Beighton or similar tests is good, but there is lack of evidence for the validity of this test assessment method [Remvig et al., 2007a].

For adults, a cut-point of 4/9 for GJH is included in the Brighton criteria for JHS, while 5/9 for GJH is the criteria for hEDS in the Villefranche criteria. For children, there is no consensus on a specific cut-point for GJH, but cut-points of 5/9, 6/9, and 7/9 have been suggested.

Further, also questionnaire assessment methods are used for classifying GJH, among which the five-part questionnaire (5PQ) [Hakim and Grahame, 2003; Mulvey et al., 2013]. The 5PQ, so far used only for adults, consists of five questions, including actual and historical information about joint hypermobility (forward bending of the back, first finger opposition, the ability to amuse friends with strange body shapes, dislocation of shoulder/knee, perception of being double-jointed). The 5PQ is claimed to have good reproducibility, in addition to satisfactory sensitivity and specificity [Hakim and Grahame, 2003]. However, clinimetric properties (reliability, different aspects of validity, and responsiveness) have not been described fully for BS, 5PQ, or other potential clinical assessment methods for classifying GJH.

Clear and valid diagnostic clinical assessment methods and criteria for classifying GJH with or without symptoms are essential, both for diagnosing JHS/hEDS and measuring treatment effects of JHS/hEDS, in children [Scheper et al., 2013] as well in adults [Palmer et al., 2014; Smith et al., 2014; Scheper et al., 2016]. In summary, there is lack of knowledge of clinimetric properties on clinical assessment methods for classifying GJH. Therefore, the purpose of this study was to perform a systematic review for identifying the clinical assessment methods for classifying GJH, to evaluate their clinimetric properties (reliability and validity), and finally to summarize the best evidence synthesis of these clinical assessment methods.

MATERIALS AND METHODS

This systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analysis statement guidelines, PRISMA, [Moher et al., 2009] and used the PICOS method to present the chosen research questions: Participants (humans with GJH and healthy controls, ranging from childhood to adults), Intervention (assessment methods for evaluation and classification of GJH), Comparison (e.g., healthy control groups), Outcomes (reliability/validity), and Study design (e.g., reliability/case control/longitudinal studies).

The overall method used in this review can be divided into four steps: (1) Compile an exhaustive list of assessment methods for GJH on the basis of an initial search (Search 1); (2) Additional search for studies including clinimetrics of the identified assessment methods from Search 1 (Search 2); (3) Critically appraisal of the methodological quality of the identified measurement properties in each study; and (4) synthesizing of the evidence as "a best evidence synthesis."

Selection Criteria

Search 1

With restrictions on the date of publication (January 01, 1965 to December 31, 2015), humans and English, the included articles had to meet the following criteria: (1) be originally published in peer-reviewed journals involving human participants; (2) include a clinical assessment method (test or questionnaire) to classify GJH; and (3) be reported in English. Studies were excluded if they: (1) contained other advanced assessment methods used as primary assessment method and not as a reference assessment; (2) were reviews, abstracts, theses, unpublished studies

("gray literature"); or (3) were animal studies.

Search 2

By using the names of the different assessment methods found in Search 1, Search 2 was initiated, and the articles were included if they: (1) explicitly outlined a purpose for evaluating clinimetric properties of an assessment method (test or questionnaire) for classifying GJH; and (2) included at least one of the clinimetric properties of reliability, validity, and responsiveness. To avoid confusion in relation to the terminology of clinimetric properties, this study relates to the COSMIN terminology, including reliability (reliability and measurement error), validity (criterion validity and hypotheses testing), and responsiveness [Schellingerhout et al., 2008; Mokkink et al., 2010].

Search Strategy and Data Sources

Search 1 (production of a list of clinical assessment methods)

The systematic review was performed by electronic and manual searches in CINAHL ($n = 153$), Embase ($n = 1{,}027$), SportDiscus ($n = 272$), and MEDLINE ($n = 833$). Furthermore, reference lists of relevant articles were hand-searched for additional literature, and the authors conferred with experts within the field of GJH, in order to make sure no relevant articles would be missing. In each of the four databases, the following search terms were used for the Electronic Search 1 for producing a method list: (joint*; hypermobility; instability; laxity; general*) AND (evaluation*; rating*; rate*; questionnaire*; test*; scale*; assess*; examin*; observ*; diagnos*; measure*) NOT (fracture*; surgical). The search terms were adjusted to the different databases where necessary. In all databases, the search fields included title, abstract, and keywords.

Search 2 (identifying clinimetric properties)

For the Electronic Search 2, using the same databases as described in Search 1, and with a total of six identified assessment methods at this point, a total of 24 searches were conducted; one in each database, on each assessment method. The following search terms were used, for retrieving studies with clinimetric properties on each of the six identified assessment methods: (psychometric*; clinimetric*; reproducibility; reliability; repeatability; responsiveness; sensitivity; specificity; validity; diagnos*; feasibility). When including the questionnaire assessment methods, the terms test* and tool* were left out. See PRISMA flow chart (Fig. 1) for the selection process.

Data Extraction and Quality Assessment

Two authors (KS/BJK) independently screened the titles and abstracts, and agreed upon a final list of assessment methods to be included in the current review. If there were any disagreements, the full paper was retrieved for detailed assessment, and consensus was achieved. A third reviewer (RHE) was included if disagreement still existed. The handling of data were performed with the use of EndNoteWeb (https://www.myendnoteweb.com/), for easy access and organizing of data. Screening for additional references were performed based on the retrieved articles.

Eligible studies for each of the retrieved assessment methods were evaluated by the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) checklist for evaluating the methodological quality on clinimetric properties—reliability, validity, and responsiveness [Mokkink et al., 2010]. The COSMIN checklist is currently the only recommended standardized method [Terwee et al., 2012], and has been used in several different studies of clinical test assessment methods [Larsen et al., 2014; Kroman et al., 2014]. The complete COSMIN checklist includes 12 boxes, covering internal consistency, reliability, measurement error, validity, and responsiveness. The current study used the reliability and validity domains, in particular box B—reliability (14 items), box F—hypothesis testing (10 items), and box H—criterion validity (7 items). After inclusion of studies, these were grouped based on the clinimetric property assessed, in reliability (intra-, inter-tester and test-retest) and validity (hypothesis testing or criterion validity). No studies on the responsiveness domain were obtained.

A compiled list for the assessment of the item "minor methodological flaws" as used previously [Larsen et al., 2014] was included (no inclusion of the target population, only for reliability domain; only one trial per measurement/lack of information on repetition; no random order of investigators/measurements; no description of any training phase; inadequate description on demographic details). For the item "other important methodological flaws" an additional list was included (inadequately described/lacking of information about subject eligibility criteria; doubt regarding the site of measurement). Final scoring of the methodological quality of each items, evaluated on a four-point scoring system, (excellent, good, fair, and poor methodological quality) was based on the "worse score counts method" in the checklist [Terwee et al., 2012].

Finally, a best evidence synthesis was performed, by compiling the assessment of the methodological quality, the actual results of the included studies, the number of studies, and the total sample size, as outlined in connection to the COSMIN evaluation [Terwee et al., 2007], and as also performed in a previous study [Kroman et al., 2014]. The rating of the best evidence synthesis ranged from strong, moderate, limited, positive/negative, conflicting, or unknown. A note was made whenever a study was rated "poor" due to only one single item. In the reliability domain this mainly concerned the item "only one measurement" (in box B), as often used in clinical examinations and always in questionnaire studies; in the validity domain studies were mainly rated "poor" due to one rating based on the item "no information on the measurement properties of the comparator instrument(s)"; in addition, a note was made to studies rated "poor" due to "subject eligibility criteria inadequately described/lacking." Studies rated "poor" due to only one "poor" item were upgraded to "fair," in line with previous systematic reviews, describing the limitations of the COSMIN when used for clinical test assessment methods [Kroman et al., 2014; Larsen et al., 2014].

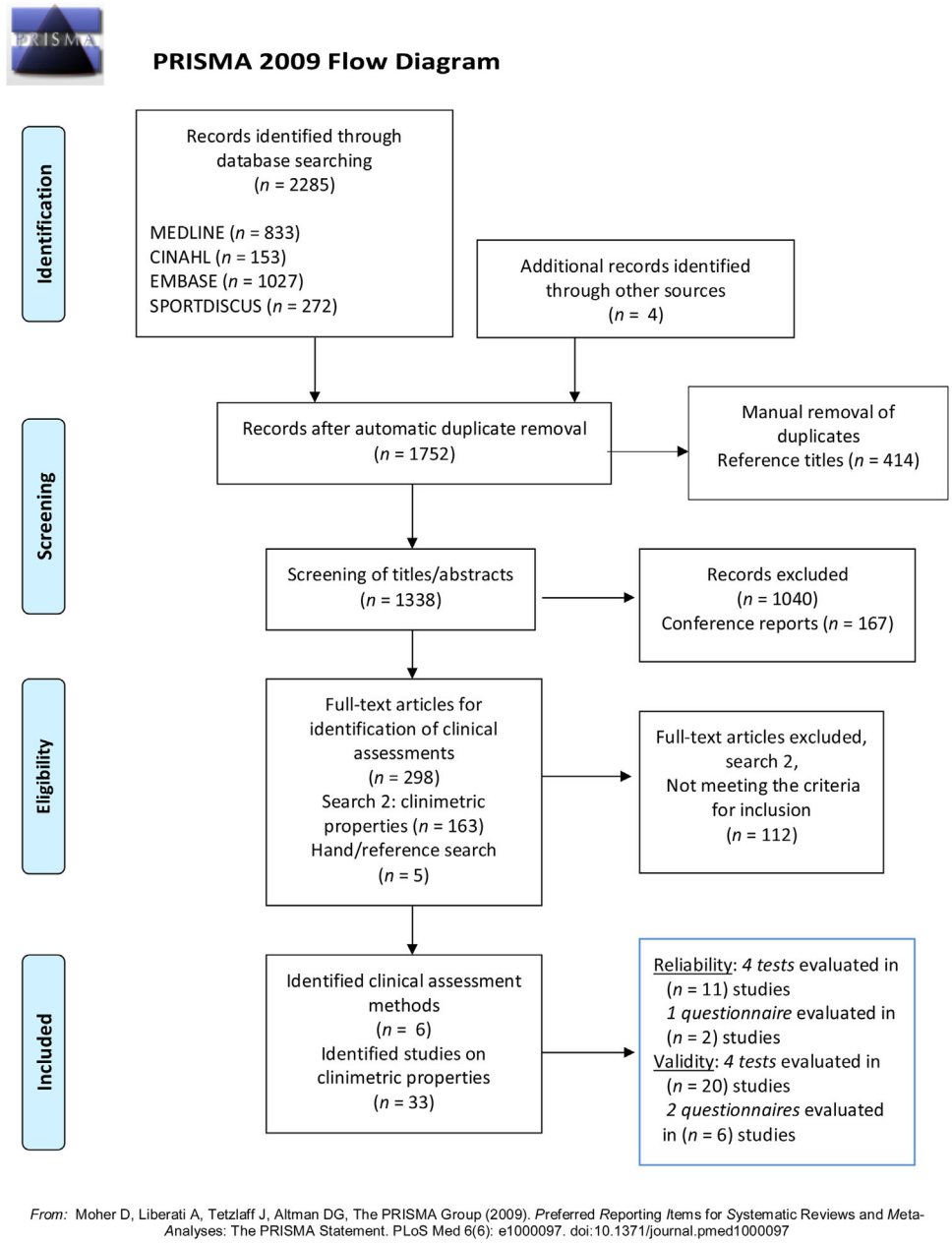

Figure 1. Flow diagram showing identification, screening, eligibility, and inclusion procedures.

Studies with more than one "poor"-rated item were omitted from the final evidence synthesis.

RESULTS

Identification of Clinical Assessment Methods

In total 2,285 references were identified, and after removal of duplicates 1,338 references were included in the screening procedure of titles and abstracts, of which 298 full-text articles were eligible according to the inclusion criteria. In Search 1, a total of six primary clinical assessment methods for classifying GJH were identified, corresponding to four test assessment methods (BS, Carter, and Wilkinson [CW], Hospital del Mar [HdM], Rotes-Querol [RQ]), and two questionnaire assessment methods (5PQ, Beighton Score self-reported [BS-self]). In Search 2, 163 references were identified, and after removal of duplicates 33 studies were identified describing the clinimetric properties of the six clinical assessment methods (Fig. 1).

Clinimetric Properties

Methodological quality in relation to reliability of the four test assessment methods (BS, CW, HdM, and RQ) was evaluated from eleven studies, and

reliability of one of the two questionnaire assessment methods (5PQ) was evaluated from two studies, while there were no reliability studies on BS-self (Table I).

Methodological quality in relation to validity of the four test assessment methods (BS, CW, HdM, RQ), was evaluated from twenty studies, and validity of the two questionnaire assessment methods (5PQ, BS-self) was evaluated from six studies (Table II).

All four tests were rated as having poor quality in all reliability studies, while 64% (7 of 11) for BS, and 50% (1 of 2) for RQ could be upgraded to fair quality, when one rating ("only one measurement") was omitted. Both studies of 5PQ were rated as having poor quality in test-retest reliability [Moraes et al., 2011; Bulbena et al., 2014], but both could be upgraded to fair. CW, RQ, and HdM were all rated as having fair quality (3/3), while for BS 82% (14/17) were upgraded and thus rated as having fair quality. All studies on validity for the two questionnaire assessment methods (for 5PQ: 5/5; for BS-self: 1/1) were upgraded, and therefore, rated as fair (Table III).

As seen in Tables I and II, all test assessment methods BS, CW, HdM, and RQ, have been tested for reliability, and for validity when compared with each other. BS has further been tested for different validity types, such as range of motion (ROM), associations with pain, injuries, and other diseases, whereas HdM has further been tested for validity on associations with shoulder injuries. 5PQ is the only one of the questionnaire assessment methods that has been tested for reliability, and for validity when compared with test assessment methods, associations with pain, diseases, and anxiety. BS-self has only been tested for validity compared with BS (Table III).

Best Evidence Synthesis: Levels of Evidence

Test assessment methods

Of the 11 reliability studies for test assessment methods, 5 studies had poor ratings on methodological quality [Bulbena et al., 1992; Mikkelsson et al., 1996; Hansen et al., 2002; Boyle et al., 2003; Aslan et al., 2006], and since they could not be upgraded to fair, they were not included in the best evidence synthesis. For the only two studies including intra-rater reliability on BS there was limited positive evidence [Erkula et al., 2005; Hirsch et al., 2007] (Table IV).

For inter-rater reliability four studies had limited positive evidence [Hicks et al., 2003; Erkula et al., 2005; Hirsch et al., 2007; Juul-Kristensen et al., 2007], while two studies had negative evidence [Karim et al., 2011; Junge et al., 2013], leaving the final evidence as limited positive to conflicting evidence. A total of four out of the eleven studies included children [Mikkelsson et al., 1996; Hansen et al., 2002; Erkula et al., 2005; Junge et al., 2013].

Validity of BS compared with other test assessment methods showed limited positive to conflicting evidence in three studies [Bulbena et al., 1992; Ferrari et al., 2005; Junge et al., 2013], while compared with ROM (trunk rotation, lower, and upper extremities) the validity in five studies showed limited negative to conflicting evidence (two on children) [Sauers et al., 2001; Erkula et al., 2005; Pearsall et al., 2006; Smits-Engelsman et al., 2011; Naal et al., 2014]. The validity for BS and the association with pain showed moderate positive to conflicting evidence in five studies (all on children) [El-Metwally et al., 2004, 2005, 2007, Tobias et al., 2013; Sohrbeck-Nøhr et al., 2014]. For the validity of BS and the association with injuries the validity showed conflicting evidence in three studies (one in children) [Rousel et al., 2009; Cameron et al., 2010; Junge et al., 2015]. For the validity of BS and the association with different diseases (Temporo-Mandibular Disorders, Chronic Fatigue Syndrome, Adhesive Capsulitis) there was limited positive to conflicting evidence in three studies [Nijs et al., 2004; Hirsch et al., 2008; Terzi et al., 2013].

CW (almost similar to the BS), and RQ had unknown evidence for both inter-rater reliability and validity compared with other test assessment methods [Bulbena et al., 1992], while HdM showed unknown evidence for inter-rater reliability and validity in the association with injuries, such as anterior shoulder dislocation [Bulbena et al., 1992; Chahal et al., 2010].

Questionnaire assessment methods

For reliability 5PQ showed conflicting evidence in the two studies [Moraes et al., 2011; Bulbena et al., 2014]. For the validity 5PQ showed limited positive to conflicting evidence compared with test assessment methods (BS, HdM) in the same two studies, while in the association with pain and tissue diseases (chronic widespread pain, JHS) 5PQ showed limited positive evidence in two studies [Hakim and Grahame, 2003; Mulvey et al., 2013], and with anxiety it showed unknown evidence [Sanches et al., 2014]. BS-self showed unknown evidence in the validity compared with BS in one study [Naal et al., 2014].

DISCUSSION

Four test assessment methods (BS, CW, HdM, RQ) and two questionnaire assessment methods (5PQ, BS-self) were identified for classifying GJH, in children and adults, with 33 studies reporting their measurement properties. The four test assessment methods and one of the questionnaire assessment methods (5PQ) reported measurement properties on both reliability and validity. Most studies were on BS, and only BS and 5PQ reported aspects of validity.

Four test assessment methods (BS, CW, HdM, RQ) and two questionnaire assessment methods (5PQ, BS-self) were identified for classifying GJH, in children and adults, with 33 studies reporting their measurement properties.

The majority of the reliability studies showed limited positive to conflicting evidence for BS, and thus, may seem acceptable to be used in clinical

TABLE I. Information From the Included Studies for COSMIN Scoring in the Reliability Domain

Ref/assessment method	Study sample/population	Registration/handling of missing data	Design	Time interval	Main result	Cut-off score	Methodological flaws
Beighton (5 items)							
Aslan et al. (2006) (BHJMI) (+goniometer for 5th finger, elbows, knees [lying])	$n=72$ Asymp. (mean age 20 yrs, range 18–25)	No information of registration/handling of missing data	Intra-/inter-rater	Intra-rater: mean 12.84 (+/− 7.41 days) Inter-rater: same day	ICC (mean) Intra: 0.92 Inter: 0.82 *% agreement* Intra: 86% (category scores) 43% (composite scores) Inter: 75%(category scores) 42% (composite scores)	Composite scores categorized (0–2: cat. 1) (3–4: cat. 2) (5–9: cat. 3)	Other important/minor methodological flaws (inadequately described/ lacking of info about subject eligibility criteria, asymp., no random order)
Boyle et al. (2003) (BHJMI) (+goniometer for 5th finger, elbows, knees [lying])	$n=42$ (intra) $n=36$ (inter) Asymp. (mean age 25 yrs, range 15–45)	No information of registration/handling of missing data	Intra-/inter-rater	Intra-rater: 1 day to 2 weeks apart Inter-rater: same day or within 6 days	*% agreement* Intra: 81% Inter: 89% Spearmans Rho (cont. score) Intra: 0.81 Inter: 0.87 ($P<.0001$) Spearmans Rho (cat. scores) Intra: .86 Inter: 0.75 ($P<.0001$)	Composite scores categorized (0–2: cat. 1) (3–4: cat. 2) (5–9: cat. 3)	Other minor methodological flaws (asymp., no random order, lacking demographic details)
Bulbena et al. (1992)	$n=30$ Symp. ($n=20$) (mean age 41 yrs) Control ($n=10$) (mean age 48 yrs)	Info on missing data provided, no info on the handling of the data	Inter-rater	No information	Kappa (range) 0.79–0.93	NS	Other minor methodological flaws (lack of info on repetition, no random order, inadequate demographic details)
Erkula et al. (2005)	$n=50$ Asymp. (children, mean age 10 yrs)	No information of registration/handling of missing data	Intra-/inter-rater	Reassessment after 2 weeks	Spearmans Rho Intra: 0.62 Inter: 0.86	≥7/9 (classified as sign. joint laxity)	Other important/minor methodological flaws (subject eligibility criteria inadequately described, asymp., no random order, no training phase, no blinding of results)

continued

TABLE I. (*Continued*)

Ref/assessment method	Study sample/population	Registration/handling of missing data	Design	Time interval	Main result	Cut-off score	Methodological flaws
Hansen et al. (2002) (4/5 selected tests included, no 5th finger)	$n=100$ Asymp. (children, 9–13 yrs)	Info on missing data provided, the handling of the data can be deduced	Inter-rater	No information	Kappa (range) 4 tests: 0.44–0.82 (experts) $\leq$0.40 (inexperienced/ parents)	NS	Other minor meth. flaws (asymp., no training phase, inadequate demographic details, no blinding)
Hicks et al. (2003)	$n=63$ Symp. (mean age 36 yrs, range 20-66)	No information of registration/handling of missing data	Inter-rater	15 min time-delay between raters	ICC (mean) Pair 1: 0.95 Pair 2: 0.76 Pair 3: 0.66	NS	Other important/minor meth. flaws (inadequately described/lacking of info about subject eligibility criteria, no random order, inadequate demographic details)
Hirsch et al. (2007) (+goniometer for elbows, knees)	$n=50$ Asymp. (mean age 38 yrs, range 20–60) (recruited from a general dental practice setting)	Info on missing data provided (1 subject), no info on the handling of the data	Intra-/inter-rater	An average of 24.6 days between first and follow-up exam	ICC (mean) Intra: >0.89 Inter: >0.84 Cronbach's alpha (intra- and inter-rater agreement) Mean 0.75 Median 0.77	$\geq$4/9	Other minor meth. flaws (asymp., no random order)
Junge et al. (2013) (2 different methods)	$n=39$ Asymp. (school children, aged 7–8 and 10–12 yrs)	No information of registration/ handling of missing data	Inter-rater	Approx. 30 min between testing sessions	% agreement 74–97% (method A) 72–97% (method B) $\geq$5/9: 82% (A), 80% (B) Kappa (range) 0.49–0.94 (method A) 0.30–0.84 (method B) $\geq$5/9: 0.64 (A), 0.59 (B)	$\geq$5/9	No other methodological flaws
Juul-Kristensen et al. (2007)	$n=40$ Symp. (mean age 34 yrs) Asymp.	No information of registration/ handling of missing data	Inter-rater	No information	Kappa (range) 0.34–1.00 (curr) 0.60–1.00 (curr/hist) $\geq$5/9: 0.66 (curr), 0.74 (curr/hist)	$\geq$5/9	No other methodological flaws

continued

TABLE I. (*Continued*)

Ref/assessment method	Study sample/population	Registration/handling of missing data	Design	Time interval	Main result	Cut-off score	Methodological flaws
	(mean age 46 yrs)				ICC (mean): 0.91(curr/hist)		
Karim et al. (2011)	$n = 30$ Contemporary Pro dancers (mean age 24 yrs, range 18–32)	No information of registration/handling of missing data	Inter-rater	No information/guideline available upon request	*% agreement* 54–100% Kappa (mean) 0.60	NS	Other minor methodological flaws (no random order, no description of blinding of results)
Mikkelsson et al. (1996)	$n = 29$ (inter) $n = 13$ (intra) Asymp. school children (mean ages 9 and 11 yrs)	No information of registration/handling of missing data	Intra-/inter-rater	Intrarater: within the same lesson (at the beginning and at the end) Inter-rater: during the same lesson	Kappa (mean) Inter: 0.78 Intra: 0.75 ICC (mean) Inter: 0.80 Intra: 0.84	≥6/9	Other minor methodological flaws (asymp., no random order, no training phase, inadequate demographic details)
Carter & Wilkinson (5 items, score 0–5)							
Bulbena et al. (1992)	$n = 30$ Symp. ($n = 20$) (mean age 41 yrs) Controls ($n = 10$) (mean age 48 yrs)	Info on missing data provided, no info on the handling of the data	Inter-rater	No information	Kappa (range) 0.68–0.92	NS	Other minor methodological flaws (lack of info on repetition, no random order, inadequate demographic details)
Hospital del Mar (9/10 items, score 0–9)							
Bulbena et al. (1992) (item testing)	$n = 30$ Symp. ($n = 20$) (mean age 41 yrs) Controls ($n = 10$) (mean age 48 yrs)	Info on missing data provided, no info on the handling of the data	Inter-rater	No information	Kappa (range) 0.61–1.00	NS	Other minor methodological flaws (lack of info on repetition, no random order, inadequate demographic details)
Rotès-Quérol (11 items, score 0–11)							
Bulbena et al. (1992)	$n = 30$ Symp. (n = 20) (mean age 41 yrs) Controls ($n = 10$) (mean age 48 yrs)	Info on missing data provided, no info on the handling of the data	Inter-rater	No information	Kappa (range) 0.44–0.93	NS	Other minor methodological flaws (lack of info on repetition, no random order, inadequate demographic details)
Juul-Kristensen et al. (2007)	$n = 40$ Symp.	No information of registration/handling of missing	Inter-rater	No information	Kappa (range) 0.32–0.79 (curr) 0.31–0.80 (curr/hist)		No other methodological flaws

continued

TABLE I. (*Continued*)

Ref/assessment method	Study sample/population	Registration/handling of missing data	Design	Time interval	Main result	Cut-off score	Methodological flaws
(3 items)	(mean age 34 yrs) Asymp. (mean age 46 yrs)	data			ICC (mean): 0.83 (curr/hist)		
5-part questionnaire (5 items, score 0–5)							
Bulbena et al. (2014)	$n = 33$	No information of registration/ handling of missing data	Test-retest	1 week	Tau-kendall index: 0.91 ICC (mean): 0.96	≥3/7	Other minor methodological flaws (no random order, inadequate description on demographic details)
5-part questionnaire (+2 questions)	Symp. (anxiety) (mean age 35 yrs)						
Moraes et al. (2011)	$n = 211$	Info on missing data implicit provided, the handling of the data can be deduced	Intra-group agreement	6 months	Kappa (mean) Q1: 0.63 Q2: 0.70 Q3: 0.65 Q4: 0.57 Q5: 0.48	≥2/5	Other minor methodological flaws (asymp., no random order, inadequate demographic details, no blinding)
5-part questionnaire	Asymp. (ages 17–24 yrs)						

BHJMI, Beighton and Horan Joint Mobility Index; NS, not stated; approx., approximately, Intra, intrarater; Inter, interrater; yrs, years; asymp., asymptomatic; symp., symptomatic; Q, question; cont., continuous; cat., category; curr, currently; hist, historically; ICC, intraclass correlation; sign., significant; rho, rank correlation co-efficient.

practice, provided that uniformity of testing procedures is included in testing procedures, in addition to historical information, especially in adults. However, there are shortcomings on studies for the validity of BS, while the three other test assessment methods lack information on both reliability and validity. For the questionnaire assessment methods, 5PQ was the most frequently used, however, only in adult population studies, and the reliability showed conflicting evidence. Concerning the validity there were shortcomings on studies for 5PQ, while for BS-self the validity showed unknown evidence in comparison with BS. More studies are needed to conclude on the measurement properties for BS-self.

Inter-rater reliability studies on test assessment methods were most frequently reported on BS, with the majority showing limited positive to conflicting evidence, and thus, may be acceptable for this assessment method. This may provide useful information for clinicians and researchers, in order to establish uniformity in carrying out the procedures. On the other hand, it also shows the need for more future comprehensive studies of this test assessment method, since unclear/vague or different descriptions of the procedures for performing the Beighton tests were used (e.g., thumbs apposition with straight or flexed elbow, knee extension in standing or supine lying). The procedures initially illustrated by photos for performing the Beighton tests [Beighton et al., 1973] are recommended for future clinical use, as described in detail in the appendix of one of the reliability studies [Juul-Kristensen et al., 2007], as they have satisfactory reliability.

Some of the studies did not include all nine tests as recommended, which especially is important when defining cut-points for classifying GJH, as discussed further below. Other test assessment methods, such as CW, RQ, and HdM showed unknown evidence on reliability in one single study [Bulbena et al., 1992], which is too limited to conclude on.

For the questionnaire assessment methods, most of the studies were

TABLE II. Information From the Included Studies for COSMIN Scoring in the Validity Domain

Ref/assessment method	Year	Study sample/population	Registration/handling of missing data	Construct validity hypotheses formulated/direction/criterion validity (adequate "gold standard")	Main results	Cut-off points	Methodological flaws (explicitly stated or lack of information)
BS and other tests/different BS procedures							
Bulbena et al. •BS •CW (5 items) •RQ (11 items)	1992	$n=114$ (Symp.) $n=59$ (controls)	Info on missing data provided, the handling of data can be deduced	Agreement between BS, CW, and RQ criteria to define the HMS Multiple hypotheses formulated a priori	Pearson correlation BS vs. RQ:0.89 BS vs. CW: 0.97 CW vs. RQ:0.86 % agreement CW ≥3: 96–98% BS ≥5: 96–98% RQ ≥5: 95–96%	Different cutpoints are tested	Other important/minor meth. flaws (subject eligibility criteria inadeq. described/ lacking for the control group, 1 trial per measure., no random order, no training phase, inadequate demographic details, no blinding)
Ferrari et al. •BS •Lower Limb Assessment Score (LLAS) (12 items)	2005	1) $n=21$ (hyp.) 2) $n=88$ (possible hyp.) 3) $n=116$ (controls) (range: 5-16 yrs)	No info on missing data/handling provided	Agreement between BS and LLAS Hypotheses vague/not formulated, but possible to deduce what was expected	% agreement BS vs. LLAS 69% (1), 80% (3) Pearson correlation 0.43 (1), 0.79 (2), 0.79 (3)	≥5/9 (BS) ≥7/12 (LLAS)	Other minor methodological flaws (1 trial per measure. session, no random order, no training phase, inadequate demographic details, no blinding)
Junge et al. •BS method A •BS method B	2013	$n=103$ Asymp. (school children, aged 7–8 and 10–12 yrs)	Info on missing data provided, the handling of the data can be deduced	Agreement between preval. of GJH of Methods A and B Hypotheses vague/not formulated, but possible to deduce what was expected	McNemar sign probability test Prevalence on Method A and B No difference ($P=0.54$)	≥5/9 (BS A + B)	No other meth. flaws
BS and ROM							
Erkula et al. •BS	2005	$n=1273$	No info on missing data/handling	Agreement between joint laxity and trunk rotation	Chi2-test BS vs. scap.	≥7/9 (BS)	Other important/minor meth. flaws

continued

TABLE II. (*Continued*)

Ref/assessment method	Year	Study sample/population	Registration/handling of missing data	Construct validity hypotheses formulated/direction/criterion validity (adequate "gold standard")	Main results	Cut-off points	Methodological flaws (explicitly stated or lack of information)
•Trunk rotation/ asymmetry		Asymp. (mean age 10.4, range 8–15 yrs)	provided	Minimal hypotheses formulated a priori	asymmetry: ($P=0.008$) BS vs. left scap. elevation: ($P=0.028$)		(inadequately described subject eligibility criteria, only 1 trial per measure. session, no random order, no training phase, inadequate demographic details, no blinding)
Naal et al. •BS •Femoroacetabular impingement (FAI)	2014	$n=55$ Symp. (mean age 28.5 yrs, SD 4.1 yrs)	No info on missing data/handling provided	Agreement between BS and Hip ROM Hypotheses vague/not formulated, but possible to deduce what was expected	Spearmans correlation BS vs. Hipflex: 0.61 Int. rot: 0.56 Ext. rot: 0.44 (all with $P<0.01$)	≥4/9 (BS) ≥6/9 (BS)	Other minor methodological flaws (only 1 trial per measurement session, no random order, no training phase, no blinding)
Pearsall et al. •BS •Knee arthrometer (KT-2000) •Ankle arthrometer	2006	$n=57$ Asymp. (mean age 20.9 yrs, SD 1.45 yrs)	No info on missing data/handling provided	Agreement between BS and knee and ankle joint specific laxity Hypotheses vague/not formulated, but possible to deduce what was expected	Spearmans correlation BS vs. Kneelax: 0.37 ($P=0.110$) A/P ankle lax: 0.21 ($P=0.152$) Int-Ext rot ankle lax: 0.24 ($P=0.101$)	NS	Other minor meth. flaws (no random order, no training phase, no blinding)
Sauers et al. •BS (4/5 items, forward flexion not included) •Shoulder	2001	$n=51/102$ Asymp. (healthy shoulders)	No info on missing data/handling provided	Agreement between BS and glenohumeral joint laxity Hypotheses vague/not formulated, but possible to deduce what was expected	Pearsons correlation Bs vs. instr. AP lax score: (0.23, NS) Clin pass ROM:	NS (BS 0–8) assumed)	Other minor methodological flaws (only 1 trial per measure. session, no random order, no

continued

TABLE II. (*Continued*)

Ref/assessment method	Year	Study sample/population	Registration/handling of missing data	Construct validity hypotheses formulated/direction/criterion validity (adequate "gold standard")	Main results	Cut-off points	Methodological flaws (explicitly stated or lack of information)
arthrometer •Clinical passive ROM		(mean age 22 yrs, SD 2.8 yrs)			(0.01–0.48, NS)		training phase, inadequate demographic details, no blinding)
Smits-Engelsman et al. •BS •Goniometer	2011	$n = 551$ Asymp. (mean age 8 yrs, range 6–12 yrs)	No info on missing data provided, but the handling of data can be deduced	Agreement between BS and 16 ROM Hypotheses vague/not formulated, but possible to deduce what was expected	Variance analysis Sign. diff in mean ROM between the 3 BS groups ($P < 0.001$, except knee flex $P = 0.02$; hip ext $P = 0.06$)	BS: 0–4 (not hyp) 5–6 (incr. hyp) 7–9(hyp)	Other minor methodological flaws (only 1 trial per measure. session, no random order, no blinding)
BS and pain							
El-Metwally et al. •BS •Musculoskeletal pain	2004	$n = 430$ Asymp. (mean age 9.8 and 11.8 yrs)	Info on missing data, and handling of data provided	Predictive baseline factors for persistence/recurrence of MSK pain, from childhood till adolescence Hypotheses vague and direction not formulated, but possible to deduce what was expected	Pain recurrence 4-yr follow-up (GLModels + RiskRatio): (RR = 1.35 [1.08–1.68])	≥6/9	Other minor methodological flaws (only 1 trial per measure, no random order, no training phase, inadequate demographic details)
El-Metwally et al. •BS •Lower limb pain (LLP)	2005	$n = 1284$ Asymp. (mean age 9.8 and 11.8 yrs)	No info on missing data provided, but the handling of data can be deduced	Predictive baseline factors for persistence/recurrence of LLP, from childhood till adolescence Hypotheses vague and direction not formulated, but possible to deduce what was expected	LLP recurrence 4-yr follow-up (GLModels + OddsRatio) (OR = 2.93 [1.13–7.70])	≥6/9	Other minor methodological flaws (only 1 trial per measure, no random order, inadequate demographic details)
El-Metwally et al. •BS •Musculoskeletal pain	2007	$n = 1113$ Asymp. (mean age 9.8 and 11.8 yrs)	Info on missing data provided, the handling of data can be deduced	Predictive baseline factors for incidenceof musculoskeletal pain, from childhood till adolescence Hypotheses vague and direction	Incidence 4-yr follow-up (GLModels + OddsRatio) BS vs. Non-traum. pain (OR:0.83 [0.44–1.56]) BS vs. Traumatic pain	≥4/9 ≥6/9	Other minor methodological flaws (only 1 trial per measure, no random order, inadequate

continued

TABLE II. (*Continued*)

Ref/assessment method	Year	Study sample/population	Registration/handling of missing data	Construct validity hypotheses formulated/direction/criterion validity (adequate "gold standard")	Main results	Cut-off points	Methodological flaws (explicitly stated or lack of information)
				not formulated, but possible to deduce what was expected	(OR: 0.70 [0.16–3.04]) NS		demographic details)
Sohrbeck-Nøhr et al. •BS •Pain (arthralgia)	2014	$n=301$ Asymp. (median age 14 yrs, range 13-15 yrs)	Info on missing data provided, described handling	Predictive baseline factors for incidence of arthralgia, from childhood till adolescence Hypotheses vague and direction not formulated, but possible to deduce what was expected	Arthralgia incidence 4 and 6-yr follow-up (Logistic Regress Models + Odds Ratio) BS ≥5/9 (OR: 3.00, [0.94–9.60])	≥4/9 ≥5/9 ≥6/9	Other minor methodological flaws (no random order)
Tobias et al. •BS •Musculoskeletal pain	2013	$n=2901$ Asymp. (mean age 13.8 yrs)	Info on missing data provided, the handling of data can be deduced	Predictive baseline factors for incidence of arthralgia, from childhood till adolescence Hypotheses vague and direction not formulated, but possible to deduce what was expected	Pain association 4-yr follow-up (Logistic Regress Models + Odds Ratio) Shoulder (OR1.68 [1.04–2.72]) Knee (OR 1.83 [1.10–3.02]) Ankle/foot (OR 1.82 [1.05–3.16])	≥6/9	Other minor methodological flaws(only 1 trial per measure, no random order, no blinding)
BS and injuries							
Cameron et al. •BS •Glenohumeral instability (GI, historic info)	2010	$n=714$ (soldiers, mean age 18.8 yrs, SD 1 yr)	Info on missing data provided, described handling	Relationship between BS and GI Hypotheses vague and direction not formulated, but possible to deduce what was expected	GI association (Logistic Regress Models + Odds Ratio) (OR 2.48 [1.19–5.20])	≥2/9	Other minor methodological flaws: (1 trial per measure, no random order)
Chahal et al. •HdM tests (10 items) + ext. rot >85° (GLL+ER) •Primary anterior shoulder dislocation (ASD)	2010	$n=57$(cases)/ 92(controls) (mean age 24 yrs, SD: 0.32)	No info on missing data provided, the handling of data can be deduced	Predictive risk factor of GLL + ER for AID Minimal number of hypotheses formulated, expected direction of correlation stated	AID association (Odds Ratio) All (OR: 2.79 [1.27–6.09]) GLL (OR: 3.6 [1.49–8.68]) GLL + ER Men: (OR: 7.43 [2.13–25.57])	Men: ≥4/10 Women: ≥5/10	Other minor methodological flaws (1 trial per measure, no random order, inadequate demographic details, no blinding)

continued

TABLE II. (*Continued*)

Ref/assessment method	Year	Study sample/population	Registration/handling of missing data	Construct validity hypotheses formulated/direction/criterion validity (adequate "gold standard")	Main results	Cut-off points	Methodological flaws (explicitly stated or lack of information)
					GLL (OR: 6.75 [1.92–23.36])		
Junge et al. •BS; Knee hypermobility (KH) •Knee injuries by SMS-track (KI)	2015	$n = 999$ (school children, 9–14 yrs)	Info on missing data provided, described handling	Predictive risk factor of BS + KH for KI Minimal number of hypotheses formulated, expected direction of correlation stated	KI association (Logistic Regress models + Odds Ratio) Traumatic Inj (OR: 1.56 [0.43–5.61]) BS TraumaticInj (OR: 2.22 [0.60–8.19]) BS + KH	≥5/9	Other minor methodological flaws (1 trial per measure, no random order, no training phase)
Roussel et al. •BS + goniometer •Pelvic injuries (PI), low back pain (LBP)	2009	$n = 32$ (dancers) $n = 26$ (students) (mean 20 yrs, SD: 2 yrs)	Info on missing data provided, described handling	Predictive risk factor of BS for PI and LBP Hypotheses vague and direction not formulated, but possible to deduce what was expected	PI/LBP (Spearman correlation: (Rho = − 0.03; $P = 0.89$) PI, LBP	BS: 0–3 (tight) 4–6 (hyp) 7–9 (extr. hyp)	Other important/minor meth. flaws(inadequately described subject eligibility criteria, 1 trial per measure, no random order, no training phase, inadequate demographic details, no blinding)
BS and diseases							
Hirsch et al. •BS •Temporo-mandibular sympt (TMs)	2008	$n = 895$ (mean age 40.6 yrs, SD: 11.6 yrs)	No info on missing data provided, but the handling of data can be deduced	Agreement between BS and TMs Hypotheses vague and direction not formulated, but possible to deduce what was expected	Association (Multiple log. Regress. Models + Odds Ratio) Joint click, reduced ROM (OR: 1.56 [1.01–2.39])	Category 0 (BS = 0) 1 (BS = 1–3) 2 (BS = 4–9)	Other minor methodological flaws (only 1 trial per measure.session, no random order, no training phase, inadequate

continued

TABLE II. (*Continued*)

Ref/assessment method	Year	Study sample/population	Registration/handling of missing data	Construct validity hypotheses formulated/direction/criterion validity (adequate "gold standard")	Main results	Cut-off points	Methodological flaws (explicitly stated or lack of information)
							demographic details, no blinding)
Nijs et al. •BS •Chronic fatigue syndrome (CFS)	2004	$n = 44$ Symp. (CFS) (mean age 40.2 yrs, SD: 9.11)	Info on missing data provided, but the handling of data can be deduced	Association between widespread pain in patients with CFS and BS Multiple hypotheses formulated a priori, direction and but magnitude not stated	Association between BS and CFS CFS: 63.6% Non-CFS: 36.5% (Fischers exact test: $P = 0.73$) CFS-APQ1: $P = 0.75$ CFS-APQ2: $P = 0.69$ (NS)	≥4/9	Other minor methodological flaws (only 1 trial per measure, no random order, no training phase, inadequate demographic details, noblinding)
Terzi et al. •BS •Adhesive capsulitis (AC) of the shoulder •Subacromial impingement syndrome (SIS)	2013	$n = 240$ (120 with AC, 120 controls/SIS) (mean age 53.9–54.08 yrs, SD: 8.21–10.68)	No info on missing data provided, but the handling of data can be deduced	Difference in GJH (BS) in AC vs. SIS Minimal number of hypotheses formulated, expected direction, but not magnitude of correlation stated	Chi-square GJH in AC vs. SIS: (0.8% vs. 7.5%, $P = 0.010$)	NS	Other minor methodological flaws (only 1 trial per measure, no random order, no training phase, no blinding)
Questionnaires (5-part) and other tests							
Bulbena et al. •5PQ (+2 questions) •HdM (10 items)	2014	$n = 191$ pt's (potential anxiety, mean age 35.5 yrs,	No info on missing data/handling provided	Agreement between 5PQ + 2Q and HDM Hypotheses vague and direction not formulated, but possible to deduce what was expected	Spearmans correlation 5PQ+2Q and HdM: (rho = 0.75; $P < 0.001$) 5PQ + 2Q and 5PQ: (rho = 0.86; $P < 0.001$)	≥3/7	Other important/minor meth. flaws (inadeq. described subject eligibility criteria, 1 trial per measure, no

continued

TABLE II. (*Continued*)

Ref/assessment method	Year	Study sample/population	Registration/handling of missing data	Construct validity hypotheses formulated/direction/criterion validity (adequate "gold standard")	Main results	Cut-off points	Methodological flaws (explicitly stated or lack of information)
		SD: 10.18 yrs)					random order, no training phase, inadeq. demographic details, no blinding)
Moraes et al. •5PQ (transl.) •BS 5-part Q and pain	2011	$n=394$ (university students, age 17–24 yrs)	Info on missing data provided, the handling can be deduced	Agreement between 5PQ and BS Hypotheses vague and direction not formulated, but possible to deduce what was expected Assumable that the criterion used (BS) can be considered a reasonable "gold standard"	% agreement 73.6% Sensitivity 70.9 (62.1–78.6) Specificity 77.4 (71.4–82.6)	≥2/5 (5PQ) ≥4/9 (BS)	Other minor methodological flaws (no random order, inadequate demographic details, no blinding)
Mulvey et al. •5-part questionnaire •Chronic widespread pain (CWP) 5-part Q and diseases	2013	$n=12{,}354$ (population, median age 55 yrs, range 25–107 yrs)	Info on missing data provided, registration/ handling described	Association between JH and musculoskeletal pain Minimal number of hypotheses formulated, expected direction of correlation stated	Relative risk ratio JH and CWP grade I: (RRR 1.2; $P=0.03$ – modest ass.) grade II: (RRR 1.2, $P=0.1$ – modest ass., not stat. sign.) grade III/IV: (RRR 1.4; $P=<0.001$)	≥2/5	Other minor methodological flaws (1 trial per measure., no random order, no training phase, inadequate demographic details, no blinding)
Hakim and Grahame •5PQ	2003	$n=269$ (212 BJHS, 57 controls) $n=220$ (170 BJHS, 50 controls) (range	No info on registration/ handling of missing data	Agreement between 5PQ and BJHS Hypotheses vague and direction not formulated, but possible to deduce what	Sens. (mean) 77-85% (1st and 2nd cohort) Spec. 80–89% (1st and 2nd cohort)	≥2/5	Other minor methodological flaws (1 trial per measure., no random order, no training phase,

continued

TABLE II. (*Continued*)

Ref/assessment method	Year	Study sample/population	Registration/handling of missing data	Construct validity hypotheses formulated/direction/criterion validity (adequate "gold standard")	Main results	Cut-off points	Methodological flaws (explicitly stated or lack of information)
		16–80 yrs)		was expected Assumable that the criterion used (BS) can be considered a reasonable "gold standard"			inadequate demographic details, no blinding)
Sanches et al. •5PQ •Anxiety	2014	$n=2300$ (university students, mean age 21 yrs, SD: 3.25)	Info on missing data provided, registration/handling can be deduced	Association between Beck Anxiety Inventory (BAI) and 5PQ Hypotheses vague and direction not formulated, but possible to deduce what was expected	Pearsons correlation 5PQ and BAI total score (Women: $r=0.11$, $P=0.007$, weak ass.) (Men: $r=0.04$, ($P=0.450$), not sign. ass.)	≥2/5	Other minor methodological flaws (1 trial per measurement, no random order, no training phase, inadequate demographic details, no blinding)
BS self-reported Naal et al. •BS •BS-self (from paper drawings) •Femoroacetabular impingement (FAI)	2014	$n=55$ Symp. (diagnosed with FAI) (mean age 28.5 yrs, SD: 4.1 yrs)	Info on missing data provided, registration/handling can be deduced	Agreement between BS-self and BS/agreement between BS and Hip ROM Hypotheses vague and direction not formulated, but possible to deduce what was expected	Kappa-values (L_w) Total 0.82 (0.72–0.91) Single tests range: 0.61–0.96 Spearmans correlation BS vs. Hipflex: 0.61 Int. rot: 0.56Ext. rot: 0.44 (all with $P<0.01$)	≥4/9 (BS) ≥6/9 (BS)	Other minor methodological flaws (1 trial per measure, no random order, no training phase, no blinding)

BS, Beighton Scale; CW, Carter & Wilkinson; HDM, Hospital del Mar; RQ, Rotès-Quérol; LLAS, Lower Limb Assessment Score; 5PQ, 5-part questionnaire; measure, measurement; meth, methodological; hyp, hypermobile; GJH, generalized joint hypermobility; HMS, hypermobility syndrome; GLL, generalized ligamentous laxity; JH, joint hypermobility; BJHS, benign joint hypermobility syndrome; MSK, musculoskeletal; scap, scapular; lax, laxity; ROM, range of motion; clin, clinical; pass, passive; incr, increased; diff, difference; sign, significant; ass, association; traum, traumatic; transl., translation; inadeq, inadequate; ER, external rotation; ASD, Primary anterior shoulder dislocation; KH, knee hypermobility; KI, knee injuries; PI, pelvic injuries; LBP, low back pain; TMs, Temporomandibular symptoms; CFS, chronic fatigue syndrome; AC, adhesive capsulitis; SIS, subacromial impingement syndrome; CWP, chronic widespread pain; BAI, Beck Anxiety Inventory; FAI, femoroacetabular impingement; A/P, anterior/posterior; SD, standard deviation; OR, odds ratio; RR, risk ratio; sens, sensitivity; spec, specificity; NS, non-significant; GLmodel, general linear regression model.

TABLE III. COSMIN Scoring of the Methodological Quality of Each Study on Measurement Properties (Reliability and Validity)

		Reliability			Validity				
Assessment methods/ reference	Study (n)/ age	Design/result	Cut-off score	COSMIN score	Reference	Study (n)/age	Design/result	Cut-off score	COSMIN score
Clinical tests									
Carter & Wilkinson (0–5) (4 bilateral BS tests + 1 ankle test)									
Bulbena et al. (1992)	$n = 30$ (mean age 41 yrs. [symp.], 48 yrs. [control])	Kappa: substantial/ almost perfect	–	Poor (only one measurement, lacking of info about subject eligibility criteria)			Hyp. test. (CW vs. RQ/BS) Pearson corr. High (CW vs. RQ) Very high (CW vs. BS) % agreement CW ≥3: 96–98%	–	Poor*** (other imp.meth. flaws: subject eligibility criteria inadeq. described/lacking for the control group)
Beighton (0–9)									
Erkula et al. (2005)	$n = 50$ (mean age 10 yrs)	Spearmans rho: Moderate (intra) High (inter)	≥7/9 (classif. as significant joint laxity)	Poor* (only one measurement)			Hyp. test. (BS vs. ROM) Chi^2-test: ($P = 0.008$) BS vs. scap. asymmetry ($P = 0.028$) BS vs left scap. elevation	≥7/9	Poor (no info on the measurement prop. of the comparator instr., other important meth. flaws - lacking of info about subject eligibility criteria)
Hirsch et al. (2007)	$n = 50$ (mean age 38 yrs, range 20–60)	ICC: Excellent (intra/inter)	–	Poor* (only one measurement)	Hirsch et al. (2008)	$n = 895$ (mean age 40.6 yrs, SD: 11.6 yrs)	Hyp. test (BS vs. TMDs) Odds Ratio: Joint click (OR: 1.56 (1.01–2.39)	Composite scores categoris.: (0: cat. 1) (1–3: cat. 2) (4–9: cat. 3)	Fair
Junge et al.	$n = 39$	Kappa:	≥5/9	Poor*	Junge et al.	$n = 109$	Hyp. test	≥5/9	Fair

continued

TABLE III. (*Continued*)

		Reliability			Validity				
Assessment methods/ reference	Study (n)/ age	Design/result	Cut-off score	COSMIN score	Reference	Study (n)/age	Design/result	Cut-off score	COSMIN score
(2013) (2 different BS methods)	(aged 7–8 and 10–12 yrs)	moderate/ substantial, substantial/ almost perfect (method A) Fair/moderate, substantial/ almost perfect (method B) ≥5/9: substantial (A), moderate (B)		(only one measurement)	(2013) Junge et al. (2015)	$n=999$ (range 9–14 yrs)	(BS vs. BS A + B) McNemar sign probability test Prevalence on Method A and B: No difference ($P=0.54$) Hyp. test. (BS vs. injuries) Odds Ratio: Traumatic inj (OR: 1.56 [0.43–5.61])	≥5/9	Poor** (inadequate info on the measurement properties of the comparator instrument)
Juul-Kristensen et al. (2007)	$n=40$ (mean age 34 yrs [symp.], 46 yrs [asymp.])	Kappa: substantial (curr), substantial (curr/hist) ICC: Excellent (curr/hist)	≥5/9	Poor* (only one measurement)	Sohrbeck-Nøhr et al. (2014)	$n=301$ (mean age 14 yrs, range 13–15 yrs)	Hyp. test. (BS and arthralgia): Odds Ratio: BS ≥5/9 (OR: 3.00, [0.94–9.60])	–	Fair
Mikkelsson et al. (1996)	$n=29$ (inter) $n=13$ (intra) (mean	Kappa: Substantial (intra/inter)	≥6/9	Poor (only 1 measure., small sample size, time interval not appropriate)	El-Metwally et al. (2004)	$n=430$ (mean age 9.8 and 11.8 yrs)	Hyp. test. (BS and pain) Risk Ratio: (RR = 1.35 [1.08–1.68])	≥6/9	Fair

continued

TABLE III. (*Continued*)

		Reliability			Validity				
Assessment methods/ reference	Study (n)/ age	Design/result	Cut-off score	COSMIN score	Reference	Study (n)/age	Design/result	Cut-off score	COSMIN score
	ages 9 and 11 yrs)				El-Metwally et al. (2005)	$n = 1{,}284$ (mean age 9.8 and 11.8 yrs)	Odds Ratio: (OR = 2.93 [1.13–7.70])	≥6/9	Fair
					El-Metwally et al. (2007)	$n = 1{,}113$ (mean age 14 yrs, range 13–15 yrs)	Odds Ratio: BS vs. non-traum. pain (OR: 0.83 [0.44–1.56]) NS Traumatic pain (OR: 0.70 [0.16–3.04]) NS	≥6/9	Fair
Aslan et al. (2006) (BHJMI)	$n = 72$ (mean age 20 yrs, range 18–25)	ICC: Excellent (intra/inter)	Composite scores categorized (0–2: cat. 1) (3–4: cat. 2) (5–9: cat. 3)	Poor (other important meth. flaws, inadequate statistics applied)		–	–	–	–
Boyle et al. (2003) (BHJMI)	$n = 36/42$ (mean age 25 yrs, range 15–45)	Spearmans rho: Excellent (intra/inter)	Composite scores categorized (0–2: cat. 1) (3–4: cat. 2) (5–9: cat. 3)	Poor (time interval not appropriate, inadequate statistics applied)		–	–	–	–
Bulbena et al. (1992)	$n = 30$ (mean age 41 yrs. [symp.], 48 yrs.	Kappa: Substantial/ almost perfect	–	Poor (only one measurement, lacking of info about subject eligibility			Hyp. test. (BS vs. CW/RQ) Pearsons corr. Very high (BS vs. CW) High (BS vs. RQ)	–	Poor*** (other imp. meth. flaws: subject eligibility criteria inadeq. described/lacking for the control group)

continued

TABLE III. (*Continued*)

		Reliability			Validity				
Assessment methods/ reference	Study (n)/ age	Design/result	Cut-off score	COSMIN score	Reference	Study (n)/age	Design/result	Cut-off score	COSMIN score
	[control])			criteria)					
Hansen et al. (2002) (4/5 tests, no 5th finger)	$n=100$ (range 9–13 yrs)	Kappa: Moderate/ substantial, substantial/ almost perfect (experts) Fair (inexp.)	–	Poor (only 1 measurem., test conditions not similar)	–	–	–	–	–
Hicks et al. (2003)	$n=63$ (mean age 36 yrs, range 20–66)	ICC: Fair/good, good/excellent	–	Poor* (only one measurement)	–	–	–	–	–
Karim et al. (2011)	$n=30$ (mean age 24 yrs, range 18–32)	Kappa: Moderate	–	Poor* (only one measurement)	–	–	–	–	–
–	–	–	–	–	Cameron et al. (2010)	$n=714$ (mean age 18.8 yrs, SD: 1 yr)	Hyp. test. (BS vs. injuries) Odds Ratio GJI (OR 2.48 [1.19– 5.20]) ($P=0.16$)	≥2/9 (95th percentile)	Fair
–	–	–	–	–	Ferrari et al. (2005)	$n=21(1)/88$ (2)/116 (3)	Hyp. test. (BS vs. other tests) % agreement BS vs. LLAS:	≥5/9	Fair

continued

TABLE III. (*Continued*)

		Reliability			Validity				
Assessment methods/ reference	Study (n)/ age	Design/result	Cut-off score	COSMIN score	Reference	Study (n)/age	Design/result	Cut-off score	COSMIN score
						(range: 5–16 yrs)	69% (1), 80% (3) Pearson correlation: Low (1), High (2), High (3)		
–	–	–	–	–	Naal et al. (2014)	$n=55$ (mean age 28.5 yrs, SD 4.1 yrs.)	Hyp. test. (BS vs. Hip ROM) Spearmans correlation BS vs. Hipflex: 0.61 Int. rot: 0.56 Ext. rot: 0.44 (all with $P<0.01$)	≥4/9 (BS) ≥6/9 (BS)	Poor* (inadequate info on the measurement prop. of comparator instr.)
–	–	–	–	–	Nijs et al. (2004)	$n=44$ (mean age 40.2 yrs, SD: 9.11)	Hyp. test. (BS vs. CFS) Association between BS and CFS CFS: 63.6% Non-CFS: 36.5% (Fischers exact test: $P=0.73$) CFS-APQ1: $P=0.75$ CFS-APQ2: $P=0.69$ (non-sign)	≥4/9	Fair
–	–	–	–	–	Pearsall et al. (2006)	$n=57$ (mean age 20.9 yrs, SD 1.45 yrs)	Hyp. test (BS vs. ROM) Spearmans correlation BS vs. Kneelax: 0.37 ($P=0.110$) AP ankle lax: 0.21, $P=0.152$ IE ankle lax: 0.24, $P=0.101$	–	Fair

continued

TABLE III. (*Continued*)

		Reliability			Validity				
Assessment methods/ reference	Study (n)/ age	Design/result	Cut-off score	COSMIN score	Reference	Study (n)/age	Design/result	Cut-off score	COSMIN score
–	–	–	–	–	Roussel et al. (2009)	$n=32/26$ (mean 20 yrs, SD: 2 yrs)	Hyp. test. (BS vs. injuries) PI/LBP (Spearman correlation: (Rho = −0.03; $P=0.89$) PI, LBP	Composite scores subgr.: (0–3: gr.1, tight) (4–6: gr.2, hyp.) (7–9: gr.3, extr. hyp.)	Poor (no info on the measurement prop. of comparator instr., other important meth. flaws - lacking of info about subject eligibility criteria)
–	–	–	–	–	Sauers et al. (2001)	$n=51/102$ (mean age 22 yrs, SD 2.8 yrs)	Hyp. test. (BS vs. ROM) Pearsons correlation BS vs. instrumented AP lax score: (0.23, Non-S) Clin pass ROM: (0.01 –0.48, Non-S)	–	Fair
–	–	–	–	–	Smits-Engelsman et al. (2011)	$n=551$ (mean age 8 yrs, range 6–12 yrs)	Hyp. test. (BS vs. ROM) Variance analysis: Sign. diff in mean ROM between the 3 BS groups ($P<0.001$, except knee flex $P=0.02$; hip ext $P=0.06$)	≥7/9	Poor (no description of the constructs measured by the comparator instr., no info on the measurement prop.)
–	–	–	–	–	Terzi et al. (2013)	$n=240$	Hyp.test. (BS vs. AC)	–	Poor** (no info on the

continued

TABLE III. (*Continued*)

		Reliability			Validity				
Assessment methods/ reference	Study (n)/ age	Design/result	Cut-off score	COSMIN score	Reference	Study (n)/age	Design/result	Cut-off score	COSMIN score
						(mean age 53.9–54.08 yrs, SD: 8.21–10.68 yrs)	Chi-square GJH in AC vs. SIS: (0.8% vs. 7.5%, $P = 0.010$)		measurement prop. of the comparator instr.)
		–	–	–	Tobias et al. (2013)	$n = 2901$ (mean age 13.8 yrs)	Hyp. test. (BS and pain) Odds Ratio: Shoulder (OR 1.68 [1.04–2.72]) Knee (OR 1.83 [1.10–3.02]) Ankle/ foot (OR 1.82 [1.05–3.16])	≥6/9	Poor** (no info on the measurement prop. of the comparator instr.)
Rotès-Quérol (0–11)									
Bulbena et al. (1992)	$n = 30$ (mean age 41 yrs. [symp.], 48 yrs. [control])	Kappa: Moderate/ substantial, substantial/ almost perfect	–	Poor (only one measurement, lacking of info about subject eligibility criteria)		Hyp. test. (RQ vs. CW/BS)	Pearson correlation High (RQ vs. BS) High (RQ vs. CW)	–	Poor*** (other imp. meth.flaws: subject eligibility criteria inadeq. described/lacking for the control group)
Juul-Kristensen et al. (2007) (3 tests)	$n = 40$ (mean age 34 yrs [symp.], 46 yrs [asymp.])	Kappa: Fair/ moderate, moderate/ substantial (curr) fair/ moderate, substantial/ almost perfect (curr/hist) ICC: Excellent	–	Poor* (only one measurement)	–	–	–	–	–

continued

TABLE III. (*Continued*)

		Reliability			Validity				
Assessment methods/ reference	Study (n)/ age	Design/result (curr/hist)	Cut-off score	COSMIN score	Reference	Study (n)/age	Design/result	Cut-off score	COSMIN score
Hospital del Mar (0–10)				Design					
Bulbena et al. (1992)	$n=30$ (mean age 41 yrs. [symp.], 48 yrs. [control])	Kappa: Substantial/ almost perfect	–	Poor (only one measurement, lacking of info about subject eligibility criteria)	–	–	–	–	–
Chahal et al. (2010)	$n=57/92$ (Mean age 24 yrs, SD: 0.32)	–	–	–		Hyp. test. (HdM and injuries)	Odds Ratio: All GLL − (OR: 2.79 [1.27–6.09]) GLL + ER − (OR:3.6 [1.49–8.68]) Men GLL − (OR: 7.43 [2.13–25.57]) GLL + ER − [OR: 6.75 (1.92–23.36])	≥4/10 (men) ≥5/10 (women)	Fair
Questionnaires 5-part questionnaire (0–5)									
Bulbena et al. (2014) (+2Q)	$n=33$ (mean age 35 yrs)	ICC: Excellent Tau-kendall index: Very high corr.	≥3/7	Poor* (only one measurement)		Hyp. test. (5PQ vs. HdM)	Spearmans correlation 5PQ+2Q and HDM: (rho = 0.75; $P<0.001$) 5PQ + 2Q and 5PQ: (rho = 0.86; $P<0.001$)	≥3/7	Poor** (no info on the measurement prop. of the comparator instr.)

continued

TABLE III. (*Continued*)

		Reliability			Validity				
Assessment methods/ reference	Study (n)/ age	Design/result	Cut-off score	COSMIN score	Reference	Study (n)/age	Design/result	Cut-off score	COSMIN score
Moraes et al. (2011)	$n = 211$ (range 17–24 yrs)	Kappa: Substantial (Q1, Q2, Q3) Moderate (Q4, Q5)	≥2/5	Poor* (only one measurement)	$n = 394$	Hyp.test. (5PQ vs. BS)	% agreement 73.6% Sens.(mean) 0.69 Spec.(mean) 0.75	≥2/5	Fair (other minor meth. flaws: inadequate description on demographic details, no blinding)
						Criterion validity			Fair (other minor)
–	–	–	–	Hakim and Grahame (2003)	$n = 212$ (range 15–80 yrs)	Hyp. test. (5PQ vs. BJHS)	Sens. (mean) 77–85% (1st and 2nd cohort) Spec. 80–89% (1st and 2nd cohort)	≥2/5	Poor** (no info on the measurement prop.of the comparator instr.)
						Criterion validity			Fair (other minor meth. flaws: inadequate description on demographic details, no description of blinding)
–	–	–	–	Mulvey et al. (2013)	$n = 2,354$ (median age 55 yrs, range 25–107 yrs)	Hyp.test. (5PQ vs. CWP)	Relative risk ratio: JH and CWP grade I: (RRR 1.2; $P = 0.03$ – modest ass.) grade II: (RRR 1.2, $P = 0.1$ – modest ass., not stat. sign.) grade III/IV: (RRR 1.4; $P = <0.001$)	≥2/5	Fair

continued

TABLE III. (*Continued*)

Assessment methods/ reference	Study (n)/ age	Reliability			Validity				
		Design/result	Cut-off score	COSMIN score	Reference	Study (n)/age	Design/result	Cut-off score	COSMIN score
–	–	–	–	Sanches et al. (2014)	$n = 2300$ (mean age 21 yrs, SD: 3.25)	Hyp.test. (5PQ vs. anxiety disease)	Pearsons correlation 5PQ and BAI total score (Women: $r = 0.11$, $P = 0.007$, weak ass) (Men: $r = 0.04$, $P = 0.450$, not sign ass)	≥2/5	Fair
Beighton self-reported (0–9)									
–	–	–	–	Naal et al. (2014)	$n = 55$ (mean age 28.5, SD: 4.1 yrs)	Hyp. test. (BS s-r vs BS)	Kappa-values (L_w) Total: 0.82 (0.72–0.91) Single tests (range): 0.61–0.96	≥4/9	Poor** (no info on the measurement prop. of the comparator instr.)

BS, Beighton Score; CW, Carter & Wilkinson; HDM, Hospital del Mar; RQ, Rotès-Quérol; LLAS, Lower Limb Assessment Score; 5PQ, 5-part questionnaire; measure, measurement; meth, methodological; prop, properties; GJH, generalized joint hypermobility, HMS, hypermobility syndrome; GLL, generalized ligamentous laxity; JH, joint hypermobility; BJHS, benign joint hypermobility syndrome; hyp, hypothesis; instr, instruments; scap, scapular; lax, laxity; ROM, range of motion; incr, increased; s-r, self-reported;p, persistent; r, recurrent; inc, incidence; classif, classified; cat, categorized; diff, difference; sign, significant; traum, traumatic; GJI, glenohumeral joint instability; ER, external rotation; ASD, Primary anterior shoulder

reported on 5PQ, which shows to be a promising assessment method for future population studies, where different domains of validity on GJH thus, can be studied more carefully. The additional questionnaire, BS-self, may also seem promising, as it contains illustrations of the test procedures for each of the BS tests. However, the questionnaire assessment methods need more evaluation before they can be used clinically, since very few studies have reported measurement properties on their reliability and validity.

The current review covers a wide range of populations, children, and adults (for adults comprising 64% [7/11] on reliability, and 62% [16/26] on validity), men and women (three studies on separately men or women), and different ethnic groups (mostly Caucasian, few on American/Canadian/Brazilian). For children, only BS has been used, while for adults, different test and questionnaire assessment methods have been used, though, still mostly BS, and 5PQ in population studies.

The current review demonstrates cut-points varying for the different clinical assessment methods. In the adult population, when using nine tests for BS, mostly one cut-point was used for classifying GJH varying between 4 and 6, but one study used 2/9 [Cameron et al., 2010]. However, also two cut-points were used, with a lower cut-point for "tight/not hypermobile" individuals varying between 1 and 4, and an upper cut-point for "hypermobile/extremely hypermobile" individuals varying between 4 and 7 [Boyle et al., 2003; Aslan et al., 2006; Hirsch et al., 2008; Roussel et al., 2009]. For the questionnaire assessment methods only one cut-point was used for classifying GJH, varying between 2 and 3 and with a different total score varying between 5 and 7 [Hakim and Grahame, 2003; Bulbena et al., 2014].

Generally, for adults, one cut-point, varying from 4 to 5 was used in BS (4/9 and 5/9), and 2/5 in 5PQ have been used. For children, one cut-point varying from 5 to 7 was used in

TABLE IV. Levels of Evidence of Included Studies

	Reliability			Validity		
	Intra/inter	Quality/(*n*)/pop		Hypothesis testing	Quality/(*n*)/pop	
Clinical assessment tools						
Beighton	Aslan[06]: ++	*P* (72)		vs. other tests:		
	Boyle[03]: ++	*P* (42/36)		(CW/RQ) Bulbena[92]: +	*P**** (173)	
	Bulbena[92]: +	*P* (30)		(LLAS) Ferrari[05]: +	*F* (225)	C
	Erkula[05]: ++	*P** (50)	C	(A/B) Junge[13]: −	*F* (103)	C
	Hansen[02]: −	*P* (100)	C	(+) to (+/−) Limited		
	Hicks[03]: +	*P** (50)		pos. to conflicting		
	Hirsch[07]: ++	*P** (50)				
	Junge[13]: −	*P** (39)	C	vs. ROM:		
	Juul-Kr[07]: +	*P** (40)		(Trunkrot.) Erkula[05]: −	*P* (1273)	C
	Karim[11]: −	*P** (30)		(Hip) Naal[14]: −	*P*** (55)	
	Mikkelss.[96]:++	*P* (29/13)	C	(k/a lax) Pearsall[06]: −	*F* (57)	
				(sh lax) Sauers[01]: −	*F* (51)	
	Ia			(gon) Smits-Eng[11]: +	*P* (551)	C
	(+) Limited pos.			(−) to (+) Limited neg. to conflicting		
	Ie					
	(+) to (+/−) Limited pos. to Conflicting			vs. pain:		
				(p/r) El-Metwally[04]: +	*F* (430)	C
				(LLP) El-Metwally[05]: +	*F* (1284)	C
				(inc) El-Metwally[07]: −	*F* (1113)	C
				(inc) Sohrb.-Nøhr[14]: +	*F* (301)	C
				(s/k/a)(inc/p)Tobias[13]:+	*P*** (2901)	C
				(++) to (+/−)		
				Moderate pos. to Conflicting		
				BS vs. injuries:		
				(GI) Cameron[10]: +	*F* (714)	
				(KI) Junge[15]: −	*P*** (999)	C
				(PI) Roussel[09]: −	*P* (58)	
				(+/−) Conflict		
				vs. diseases:		
				(TMs) Hirsch[08]: +	*F* (895)	
				(CFS) Nijs[04]: −	*F* (44)	
				(ACprev) Terzi[13]: +	*P*** (240)	
				(+)to (+/−) Limited pos. to conflicting		
Carter & Wilkinson	Bulbena[92]: +	*P* (30)		vs. other tests:		
	Inter: Unknown			(BS/RQ) Bulbena[92]: +	*P**** (173)	
				Unknown		
Rotès-Quérol	Bulbena[92]: +	*P* (30)		vs. other tests:		
	Juul-Kr[07]: + (3)	*P**(40)		(CW/BS) Bulbena[92]: +	*P**** (173)	
	Inter: Unknown			Unknown		
Hospital del Mar	Bulbena[92]: +	*P* (30)		(ASD) Chahal[10]: +	*F* (149)	
	Inter: Unknown			Unknown		

continued

TABLE IV. (*Continued*)

	Reliability		Validity	
	Intra/inter	Quality/(*n*)/pop	Hypothesis testing	Quality/(*n*)/pop
Questionnaires				
5-part Q.	Bulbena[14]: + Moraes[11]: − Retest: (+/−) Conflicting	*P** (33) *P** (211)	vs. other tests: (HdM) Bulbena[14]: + (BS) deMoraes[11]: −/+ (+) to (+/−) Limited pos. to Conflicting	*P*** (191) *F* (394)
	Criterion val. (BS) deMoraes[11]: − (BS) Hakim[03]:? Unknown	 *F* (394) *F* (489)	vs. pain/tissue diseases: (CWP) Mulvey[13]: + (BJHS) Hakim[03]: + (+) Limited pos.	*F* (2354) *P*** (489)
			vs. anxiety/psych. disease: (anx) Sanches[14]: − Unknown	*F* (2300)
BS-self-reported			vs. tests: (BS) Naal[14]: + Unknown	*P*** (55)

Ia, Intra-rater; Ie, Inter-rater; P, poor; F, fair; pop, population; BS, Beighton Score; CW, Carter & Wilkinson; HdM, Hospital del Mar; RQ, Rotès-Quérol; 5-part Q, 5-part questionnaire; LLAS, Lower Limb Assessment Score; k/a lax, Knee/ankle laxity; sh lax, Shoulder laxity; gon, goniometry; p, persistent; r, recurrent; inc, incidence; s/k/a, shoulder/knee/ankle; GI, Glenohumeral joint instability; Y, youth; C, children, AID, Primary traumatic anterior shoulder dislocation; KI, knee injuries; PI, pelvic injuries; TMs, Temporomandibular symptoms; CFS, chronic fatigue syndrome; ACprev, Adhesive capsulitis prevalence; ASD, Anterior Shoulder Dislocation; CWP, chronic widespread pain; BJHS, benign joint hypermobility syndrome; Anx, anxiety; [92], Superscripts shows year of publication.

*Reliability studies rated poor on the basis of one single rating, based on "only one measurement."

**Validity studies rated poor on the basis of one single rating "no information on the measurement properties of the comparator instrument(s)."

BS (5/9, 6/9, and 7/9). In one test assessment method the total score was 11 with a cut-point of 7/11 for classifying GJH based on only tests in the lower extremities [Ferrari et al., 2005], and another study used two cut-points the lower being 5/9 and the upper being 7/9 [Smits-Engelsman et al., 2011]. Since JHS and hEDS in the current review are recognized as one and the same condition, a specific cut-point needs to be decided, and 5/9 may be suggested for future use in adults. However, since joint mobility, and therefore, BS is known to decrease by age [Remvig et al., 2007b], there is a need for adults also to include additional historical information, as described in the appendix of the reliability study using the BS, with phrasing "can you now or have you previously been able to ... " [Juul-Kristensen et al., 2007], and in the study describing 5PQ [Hakim and Grahame, 2003].

Generally, for adults, one cut-point, varying from 4 to 5 was used in BS (4/9 and 5/9), and 2/5 in 5PQ have been used. For children, one cut-point varying from 5 to 7 was used in BS (5/9, 6/9, and 7/9).

Since children have individual growth periods, this may be the reason for using two cut-points (a lower and an upper) as recently suggested [Smits-Engelsman et al., 2011], and therefore, the upper cut-point is suggested to be at least 6/9 as used in previous population studies [El-Metwally et al., 2004, 2005, 2007; Tobias et al., 2013].

Warming-up before performing flexibility tests may influence the outcome of a test assessment method. However, almost no studies reported whether participants did warm-up, and

the influence of such performance is therefore, unknown.

This review highlights a number of areas warranting future research. Because of the limited studies on the clinical assessment methods for classifying GJH, more high quality studies, and especially those evaluating aspects of validity are required (concurrent, predictive, measurement error, responsiveness, and interpretability). Additional clinical test assessment methods may further be considered in order to support and endorse the presence of GJH in the diagnostic procedure of heritable connective tissue disorders. Also of importance is that consensus is warranted regarding selection of specific test and questionnaire assessment methods for classifying GJH, the test performance, and the cut-points by which age, gender, and ethnicity may be taken into account.

Limitations of the study are the small amount of studies, for which reason it was decided only to rate reliability (intra- and inter-rater) and validity (hypothesis testing or criterion validity). Use of COSMIN is recommended to be the best evaluation method until now, as has also been used previously to evaluate clinical test assessment methods [Kroman et al., 2014; Larsen et al., 2014]. However, since the COSMIN originally was designed for the evaluation of patient-reported outcomes, there are some adjustments that need to be considered when using COSMIN for clinical test assessment methods. For example, although rating of the number of measurements taken may be useful in settings with continuous scales as in performance-based methods, rating scales in many clinical assessment methods (test or questionnaire) are dichotomous (positive/negative). To adjust for this shortcoming, the present review adjusted the evaluation of methodological quality, corresponding to when "only one measurement" was rated poor in reliability, the study was upgraded from poor to fair, meaning that the study thereby could be included in the best evidence synthesis. Furthermore, the sample size in clinical studies is often much smaller than in questionnaire studies, and therefore, it may be suggested that minor sample sizes should not be rated that strictly as in questionnaire assessment methods, when studying clinical test assessment methods. For validity studies, one poor rating including "no information on the measurement properties of the comparator instrument" or "subject eligibility criteria inadequately described/lacking," allowed upgrading to fair, and the study could thereby be included in the best evidence synthesis.

Strengths of this review are the systematic and rigid use of recommended strategies for systematic reviews of clinical assessment methods, the evaluation of their clinimetric properties and rating of the best evidence synthesis.

CONCLUSION

In the current review, four test and two questionnaire assessment methods for classifying GJH were found with measurement properties of varying methodological strength and results of varying weight. Most of the studies used the BS. The inter-rater reliability of this method seems acceptable to be used in clinical practice, provided that uniformity of testing procedures are included in the testing procedures, in addition to historical information, especially in adults. However, shortcomings were found in studies on the validity of BS, while the three other test assessment methods (CW, RQ, HdM) lack satisfactory information on both reliability and validity. Regarding questionnaire assessment methods, 5PQ is the most frequently method used, however, only in adult population studies. In conclusion provided uniformity of testing procedures, the recommendation for clinical use in adults is BS with cut-point of 5 of 9 including historical information, while in children it is BS with cut-point of at least 6 of 9. However, more studies are needed to conclude, especially on the validity properties of these assessment methods, and before evidence-based recommendations can be made for clinical use on the "best" assessment method for classifying GJH.

In the current review, four test and two questionnaire assessment methods for classifying GJH were found with measurement properties of varying methodological strength and results of varying weight. Most of the studies used the BS.

AUTHORS' CONTRIBUTIONS

BJK contributed to conception and design of the study, including analysis and interpretation of data, writing of the article, critical revision of the article for important intellectual content, and final approval of the article. KS contributed to the collection and assembly of data, analysis and interpretation of data, and critical revision of the article for important intellectual content and final approval of the article. RE and HL contributed to conception and design of the study including analysis and interpretation of data, critical revision of the article for important intellectual content, and final approval of the article. LR contributed to interpretation of data, critically revision of the article for important intellectual content, and final approval of the article. First and last authors take responsibility for the integrity of the work as a whole, from inception to finished article.

ACKNOWLEDGMENT

Dr. Jaime Bravo, Chile, and Hylke Bosma, previous patient representative of the Ehlers–Danlos Syndrome organization, the Netherlands, are acknowledged for their initial participation in this work and the Beighton group.

REFERENCES

Aslan BA, Celik E, Cavlak U, Akdag B. 2006. Evaluation of inter-rater and intra-rater reliability of BEighton and Horan Joint Mobility Index. Fizyoterapi Rehabil 17:113–119.

Beighton P, De Paepe A, Steinmann B, Tsipouras P, Wenstrup RJ. 1998. Ehlers–Danlos syndromes: Revised nosology, villefranche, 1997. Ehlers–Danlos national foundation (USA) and Ehlers–Danlos support group (UK). Am J Med Genet 77:31–37.

Beighton P, Solomon L, Soskolne CL. 1973. Articular mobility in an African population. Annl Rheum Dis 32:413–418.

Boyle KL, Witt P, Riegger-Krugh C. 2003. Intra-rater and inter-rater reliability of the Beighton and Horan Joint Mobility Index. J Athl Train 38:281–285.

Bulbena A, Duro JC, Porta M, Faus S, Vallescar R, Martin-Santos R. 1992. Clinical assessment of hypermobility of joints: Assembling criteria. J Rheumatol 19:115–122.

Bulbena A, Mallorqui-Bague N, Pailhez G, Rosado S, Gonzalez I, Blanch-Rubio J, Carbonell J. 2014. Self-reported screening questionnaire for the assessment of Joint Hypermobility Syndrome (SQ-CH), a collagen condition, in Spanish population. Eur J Psychiat 28:17–26.

Cameron KL, Duffey ML, DeBerardino TM, Stoneman PD, Jones CJ, Owens BD. 2010. Association of generalized joint hypermobility with a history of glenohumeral joint instability. J Athl Train 45:253–258.

Castori M, Dordoni C, Valiante M, Sperduti I, Ritelli M, Morlino S, Chiarelli N, Celletti C, Venturini M, Camerota F, Calzavara-Pinton P, Grammatico P, Colombi M. 2014. Nosology and inheritance pattern(s) of joint hypermobility syndrome and Ehlers–Danlos syndrome, hypermobility type: A study of intrafamilial and interfamilial variability in 23 Italian pedigrees. Am J Med Genet Part A 164A:3010–3020.

Chahal J, Leiter J, McKee MD, Whelan DB. 2010. Generalized ligamentous laxity as a predisposing factor for primary traumatic anterior shoulder dislocation. J Shoulder Elbow Surg 19:1238–1242.

El-Metwally A, Salminen JJ, Auvinen A, Kautiainen H, Mikkelsson M. 2004. Prognosis of non-specific musculoskeletal pain in preadolescents: A prospective 4-year follow-up study till adolescence. Pain 110:550–559.

El-Metwally A, Salminen JJ, Auvinen A, Kautiainen H, Mikkelsson M. 2005. Lower limb pain in a preadolescent population: Prognosis and risk factors for chronicity—A prospective 1- and 4-year follow-up study. Pediatrics 116:673–681.

El-Metwally A, Salminen JJ, Auvinen A, Macfarlane G, Mikkelsson M. 2007. Risk factors for development of non-specific musculoskeletal pain in preteens and early adolescents: A prospective 1-year follow-up study. BMC Musculoskelet Disord 8:46.

Erkula G, Kiter AE, Kilic BA, Er E, Demirkan F, Sponseller PD. 2005. The relation of joint laxity and trunk rotation. J Pediatr Orthop Part B 14:38–41.

Ferrari J, Parslow C, Lim E, Hayward A. 2005. Joint hypermobility: The use of a new assessment tool to measure lower limb hypermobility. Clin Exp Rheumatol 23: 413–420.

Fleiss JL. 1986. Reliability of measurement. The design and analysis of clinical experiments. New York: John Wiley and Sons. pp. 1–32.

Grahame R, Bird HA, Child A. 2000. The revised (Brighton 1998) criteria for the diagnosis of benign joint hypermobility syndrome (BJHS). J Rheumatol 27:1777–1779.

Hakim AJ, Grahame R. 2003. A simple questionnaire to detect hypermobility: An adjunct to the assessment of patients with diffuse musculoskeletal pain. Int J Clin Pract 57: 163–166.

Hansen A, Damsgaard R, Kristensen JH, Baggers J, Remvig L. 2002. Interexaminer reliability of selected tests for hypermobility. J Orthop Med 25:48–51.

Hicks GE, Fritz JM, Delitto A, Mishock J. 2003. Interrater reliability of clinical examination measures for identification of lumbar segmental instability. Arch Phys Med Rehab 84:1858–1864.

Hirsch C, Hirsch M, John MT, Bock JJ. 2007. Reliability of the Beighton Hypermobility Index to determinate the general joint laxity performed by dentists. J Orof Orthop 68:342–352.

Hirsch C, John MT, Stang A. 2008. Association between generalized joint hypermobility and signs and diagnoses of temporomandibular disorders. Eur J Oral Sci 116:525–530.

Jansson A, Saartok T, Werner S, Renstrom P. 2004. General joint laxity in 1845 Swedish school children of different ages: Age- and gender-specific distributions. Acta Paediatr 93: 1202–1206.

Junge T, Jespersen E, Wedderkopp N, Juul-Kristensen B. 2013. 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 13:214.

Junge T, Runge L, Juul-Kristensen B, Wedderkopp N. 2015. The extent and risk of knee injuries in children aged 9–14 with Generalised Joint Hypermobility and knee joint hypermobility—The CHAMPS-study Denmark. BMC Musculoskelet Disord 15:1–11.

Juul-Kristensen B, Rogind H, Jensen DV, Remvig L. 2007. Inter-examiner reproducibility of tests and criteria for generalized joint hypermobility and benign joint hypermobility syndrome. Rheumatology 46:1835–1841.

Karim A, Millet V, Massie K, Olson S, Morganthaler A. 2011. Inter-rater reliability of a musculoskeletal screen as administered to female professional contemporary dancers. Work 40:281–288.

Kroman SL, Roos EM, Bennell KL, Hinman RS, Dobson F. 2014. Measurement properties of performance-based outcome measures to assess physical function in young and middle-aged people known to be at high risk of hip and/or knee osteoarthritis: A systematic review. Osteoarthr Cartil 22:26–39.

Larsen CM, Juul-Kristensen B, Lund H, Sogaard K. 2014. Measurement properties of existing clinical assessment methods evaluating scapular positioning and function. A systematic review. Physiother Theory Pract 30:453–482.

Landis JR, Koch GG. 1977. The measurement of observer agreement for categorical data. Biometrics 33:159–174.

Mikkelsson M, Salminen JJ, Kautiainen H. 1996. Joint hypermobility is not a contributing factor to musculoskeletal pain in preadolescents. J Rheumatol 23:1963–1967.

Moher D, Liberati A, Tetzlaff J, Altman DG. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. J Clin Epidemiol 62:1006–1012.

Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, Bouter LM, de Vet HC. 2010. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. BMC Med Res Methodol 10:22.

Moraes DA, Baptista CA, Crippa JA, Louzada-Junior P. 2011. Translation into Brazilian Portuguese and validation of the five-part questionnaire for identifying hypermobility. Rev Bras Reumatol 51:53–69.

Mulvey MR, Macfarlane GJ, Beasley M, Symmons DP, Lovell K, Keeley P, Woby S, McBeth J. 2013. Modest association of joint hypermobility with disabling and limiting musculoskeletal pain: Results from a large-scale general population-based survey. Arthritis Care Res (Hoboken) 65:1325–1333.

Naal FD, Hatzung G, Muller A, Impellizzeri F, Leunig M. 2014. Validation of a self-reported Beighton score to assess hypermobility in patients with femoroacetabular impingement. Int Orthop 38:2245–2250.

Nijs J, Meirleir KD, Truyen S. 2004. Hypermobility in patients with chronic fatigue syndrome: Preliminary observations. J Musculoskelet Pain 12:9–17.

Palmer S, Bailey S, Barker L, Barney L, Elliott A. 2014. The effectiveness of therapeutic exercise for joint hypermobility syndrome: A systematic review. Physiotherapy 100:220–227.

Pearsall AW, Kovaleski JE, Heitman RJ, Gurchiek LR, Hollis JM. 2006. The relationships between instrumented measurements of ankle and knee ligamentous laxity and generalized joint laxity. J Sports Med Phys Fitness 46:104–110.

Remvig L, Engelbert RH, Berglund B, Bulbena A, Byers PH, Grahame R, Juul-Kristensen B, Lindgren KA, Uitto J, Wekre LL. 2011. Need for a consensus on the methods by which to measure joint mobility and the definition of norms for hypermobility that reflect age, gender and ethnic-dependent variation: Is revision of criteria for joint hypermobility syndrome and Ehlers–Danlos syndrome hypermobility type indicated? Rheumatology 50:1169–1171.

Remvig L, Jensen DV, Ward RC. 2007a. Are diagnostic criteria for general joint hypermobility and benign joint hypermobility syndrome based on reproducible and valid tests? A review of the literature. J Rheumatol 34:798–803.

Remvig L, Jensen DV, Ward RC. 2007b. Epidemiology of general joint hypermobility and basis for the proposed criteria for benign joint hypermobility syndrome: Review of the literature. J Rheumatol 34:804–809.

Rombaut L, Malfait F, Cools A, De Paepe A, Calders P. 2010. Musculoskeletal complaints, physical activity and health-related quality of life among patients with the Ehlers–Danlos syndrome hypermobility type. Disabil Rehabil 32:1339–1345.

Roussel NA, Nijs J, Mottram S, Van Moorsel A, Truijen S, Stassijns G. 2009. Altered lumbopelvic movement control but not generalized joint hypermobility is associated with increased injury in dancers. A prospective study. Man Ther 14:630–635.

Sanches SB, Osorio FL, Louzada-Junior P, Moraes D, Crippa JA, Martin-Santos R. 2014. Association between joint hypermobility and anxiety in Brazilian university students: Gender-related differences. J Psychosom Res 77:558–561.

Sauers EL, Borsa PA, Herling DE, Stanley RD. 2001. Instrumented measurement of glenohumeral joint laxity and its relationship to passive range of motion and generalized joint laxity. Am J Sports Med 29:143–150.

Schellingerhout JM, Verhagen AP, Thomas S, Koes BW. 2008. Lack of uniformity in diagnostic labeling of shoulder pain: Time for a different approach. Man Ther 13:478–483.

Scheper MC, Engelbert RH, Rameckers EA, Verbunt J, Remvig L, Juul-Kristensen B. 2013. Children with generalised joint hypermobility and musculoskeletal complaints: State of the art on diagnostics, clinical characteristics, and treatment. Biomed Res Int 2013:121054.

Scheper MC, Juul-Kristensen B, Rombaut L, Rameckers EA, Verbunt J, Engelbert RH. 2016. Disability in adolescents and adults diagnosed with hypermobility related disorders: A meta-analysis. Arch Phys Med Rehabil 97:2174–2187.

Smith TO, Bacon H, Jerman E, Easton V, Armon K, Poland F, Macgregor AJ. 2014. Physiotherapy and occupational therapy interventions for people with benign joint hypermobility syndrome: A systematic review of clinical trials. Disabil Rehabil 36:797–803.

Smits-Engelsman B, Klerks M, Kirby A. 2011. Beighton score: A valid measure for generalized hypermobility in children. J Pediatr 158:119–123,23 e1–4.

Sohrbeck-Nøhr O, Kristensen J, Boyle E, Remvig L, Juul-Kristensen B. 2014. Generalized joint hypermobility in childhood is a possible risk for the development of joint pain in adolescence: A cohort study. BMC Pediatr 14:1–9.

Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC. 2007. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 60:34–42.

Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. 2012. Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. Qual Life Res 21:651–657.

Terzi Y, Akgun K, Aktas I, Palamar D, Can G. 2013. The relationship between generalised joint hypermobility and adhesive capsulitis of the shoulder. Turk J Rheumatol 28:234–241.

Tinkle BT, Bird HA, Grahame R, Lavallee M, Levy HP, Sillence D. 2009. The lack of clinical distinction between the hypermobility type of Ehlers–Danlos syndrome and the joint hypermobility syndrome (a.k.a. hypermobility syndrome). Am J Med Genet Part A 149A: 2368–2370.

Tobias JH, Deere K, Palmer S, Clark EM, Clinch J. 2013. Hypermobility is a risk factor for musculoskeletal pain in adolescence: Findings from a prospective cohort study. Arthritis Rheuma 65:1107–1115.