Measurement Properties of Visual Analogue Scale, Numeric Rating Scale, and Pain Severity Subscale of the Brief Pain Inventory in Patients With Low Back Pain

Chiarotto, Alessandro; Maxwell, Lara J.; Ostelo, Raymond W.; Boers, Maarten; Tugwell, Peter; Terwee, Caroline B.

published in
Journal of Pain
2019

DOI (link to publisher)
10.1016/j.jpain.2018.07.009

document version
Publisher's PDF, also known as Version of record

document license
Article 25fa Dutch Copyright Act

The Journal of Pain, Vol 20, No 3 (March), 2019: pp 245−263
Available online at www.jpain.org and www.sciencedirect.com

Critical Review

Measurement Properties of Visual Analogue Scale, Numeric Rating Scale, and Pain Severity Subscale of the Brief Pain Inventory in Patients With Low Back Pain: A Systematic Review

Alessandro Chiarotto,[*,†] Lara J. Maxwell,[‡] Raymond W. Ostelo,[*,†] Maarten Boers,[*,§] Peter Tugwell,[‡,¶] and Caroline B. Terwee[*]

[*]Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, VU University Medical Center, Amsterdam, Netherlands.
[†]Department of Health Sciences, Amsterdam Movement Sciences Research Institute, Vrije Universiteit, Amsterdam, Netherlands
[‡]Centre for Practice-Changing Research, Ottawa Hospital Research Institute, University of Ottawa, Ottawa, Canada.
[§]Amsterdam Rheumatology and Immunology Center, VU University Medical Center, Amsterdam, Netherlands.
[¶]Department of Medicine, Faculty of Medicine, University of Ottawa, Ottawa, Canada.

Abstract: The Visual Analogue Scale (VAS), Numeric Rating Scale (NRS), and Pain Severity subscale of the Brief Pain Inventory (BPI-PS) are the most frequently used instruments to measure pain intensity in low back pain. However, their measurement properties in this population have not been reviewed systematically. The goal of this study was to provide such systematic evidence synthesis. Six electronic sources (MEDLINE, EMBASE, CINAHL, PsycINFO, SportDiscus, Google Scholar) were searched (July 2017). Studies assessing any measurement property in patients with nonspecific low back pain were included. Two reviewers independently screened articles and assessed risk of bias using the COSMIN checklist. For each measurement property, evidence quality was rated as high, moderate, low, or very low (GRADE approach) and results were classified as sufficient, insufficient, or inconsistent. Ten studies assessed the VAS, 13 the NRS, 4 the BPI-PS. The 3 instruments displayed low or very low quality evidence for content validity. High-quality evidence was only available for NRS insufficient measurement error. Moderate evidence was available for NRS inconsistent responsiveness, BPI-PS sufficient structural validity and internal consistency, and BPI-PS inconsistent construct validity. All VAS measurement properties were underpinned by no, low, or very low quality evidence; likewise, the other measurement properties of NRS and BPI-PS.

Perspectives:* *Despite their broad use, there is no evidence clearly suggesting that one among VAS, NRS, and BPI-PS has superior measurement properties in low back pain. Future adequate quality head-to-head comparisons are needed and priority should be given to assessing content validity, test-retest reliability, measurement error, and responsiveness.

Key words: *Low back pain, pain intensity, visual analogue scale, numeric rating scale, Brief Pain Inventory.*

Supported by the Task Force for Research of the Spine Society of Europe (EUROSPINE) [grant number: EUROSPINE TFR 5-2015]. These funding bodies did not have any role in in designing the study, in collecting, analyzing and interpreting the data, in writing this manuscript, and in deciding to submit it for publication.
The authors have no conflicts of interest to declare.

Supplementary data accompanying this article are available online at www.jpain.org and www.sciencedirect.com.

Address reprint requests to Alessandro Chiarotto, Department of Epidemiology and Biostatistics, Amsterdam Public Health research institute, Amsterdam Movement Sciences research institute, VU University Medical Center, de Boelelaan 1089a, Medical Faculty F-vleugel, 1081HV, Amsterdam, the Netherlands. E-mail: a.chiarotto@vumc.nl

1526-5900/$36.00

https://doi.org/10.1016/j.jpain.2018.07.009

Low back pain (LBP) is the most disabling health condition worldwide.[33] Measuring the impact of LBP on patients' lives is fundamental to monitoring clinical management and to study the (cost) effectiveness of treatments.[4] Patients with LBP have indicated that the most important domains to be measured are physical functional activities, pain reduction, quality of life, enjoyment of life, emotional well-being, and fatigue.[9,43,103] A core outcome set initiative (involving patients) aimed at standardizing measurement for LBP identified 4 core outcome domains for clinical trials: physical functioning, pain intensity, health-related quality of life, and number of deaths.[9] Among these domains, pain intensity is the most frequently assessed in LBP clinical trials.[31]

Pain intensity, defined as "how much a patient hurts, reflecting the overall magnitude of the pain experience,"[102] is the pain domain that ranked the highest among various pain domains (eg, pain quality, temporal aspects of pain, pain behavior, and pain interference) in consensus exercises to establish core outcome domains for LBP[9] and other pain conditions.[53,77] The visual analogue scale (VAS) is the patient-reported outcome measure (PROM) most frequently used to measure pain intensity in LBP trials, followed by the numeric rating scale (NRS) and the Pain Severity subscale of the Brief Pain Inventory (BPI-PS).[7,31] Recent consensus-based studies have shown that researchers and clinicians prefer the NRS over other instruments to measure pain intensity in LBP.[13,17,22,24] However, this choice has not been explicitly based on its measurement properties and feasibility.[5,85]

The NRS, VAS, and BPI-PS are highly feasible for clinical research and practice, providing very little burden to professionals and patients.[39] Various reviews have attempted to synthesize their measurement properties in samples of patients with pain.[6,42,47,52,86,94,108] All these reviews focused on chronic pain broadly and two of them solely focused in children and adolescents.[6,94] In recent years, the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) initiative has developed tools that allow researchers to conduct high quality systematic reviews on the measurement properties of PROMs.[72,73,84,100] Given that these existing reviews predated the COSMIN guidance ,[42,47,52,86,108] key methodologic steps (eg, quality assessment of the studies, formulation of evidence synthesis and findings taking the quality of the studies into account, definition of the methods to combine study results[90,105]) could not be included. Therefore, it is timely to adopt the most recent methodologic advancements in a systematic review on PROMs for pain intensity.

The objective of this study was to systematically synthesize the evidence on the measurement properties of the VAS, NRS, and BPI-PS in adult patients with LBP. This review was conducted within an international collaboration aimed at developing a core outcome measurement set for LBP[12] and informed a Delphi study to reach consensus on which core outcome measurement instrument(s) to endorse for pain intensity in LBP clinical trials.[8] For this reason, in contrast with previous reviews that had a more generic focus on various pain conditions,[6,42,47,52,86,94,108] this review focused solely on studies in patients with LBP, following the approach adopted in Cochrane reviews of randomized clinical trials on the effectiveness of interventions in patients with LBP.[32]

Methods

This systematic review was conducted according to COSMIN guidance[84] and reported according to the PRISMA statement.[71] Its protocol was registered in the international prospective register of systematic reviews (http://www.crd.york.ac.uk/PROSPERO/), registration number: CRD42015020006.

Measurement Instruments

The VAS is a self-reported scale consisting of a horizontal or vertical line, usually 10 cm long (100 mm) anchored at the extremes by 2 verbal descriptors referring to the pain status.[45] An introductory question (with or without a time recall period) asks the patient to tick the line on the point that best refers to his or her pain. The introductory question, the recall period, and the content of the external verbal descriptors vary in the literature.[39]

The NRS is a numbered version of the VAS in which the patient can select one number that best describes the pain.[23] Like in the VAS, the NRS introductory question, time recall period and verbal descriptors can vary; the most frequently used version is the 11-point (0-10) NRS.[39]

The BPI-PS consists of four 11-point NRSs, two of which asking the patient to rate the pain at its worst and least in the last 24 hours, and the other two asking about pain on the average and right now.[15] For each NRS, the verbal descriptors are no pain and pain as bad as you can imagine, and this questionnaire is usually administered as part of the BPI, which includes other 11 pain-related questions (seven of which belonging to the pain interference subscale).[15]

Literature Search

Data Sources and Searches

MEDLINE (through the interface PubMed), EMBASE (Embase.com), CINAHL (EBSCOhost), PsycINFO (EBSCOhost), and SportDiscus (EBSCOhost) were last searched on July 25, 2017. The search strategy consisted of 3 groups of search terms combined with the Boolean operator AND 1) PROMs names, 2) LBP, 3) measurement properties. A validated search filter for retrieving studies on measurement properties in PubMed was used[98]; the same filter was adapted for all the other databases (Appendix 1). No restrictions for language or time were adopted in the search strategies. Google Scholar was also searched (last on July 28, 2017) with the full names of the PROMs and the first 100 hits for each PROM were screened for inclusion. Citation tracking of the eligible studies was carried out by consulting the database Web of Science and by checking their references.

Study Selection

Any study on 1 or more of the 3 instruments was included if it assessed ≥1 of the 9 measurement properties identified by the COSMIN taxonomy: internal consistency, test-retest reliability, measurement error, content validity, structural validity, construct validity/hypotheses testing, cross-cultural validity, criterion validity, and responsiveness.[73] Studies presenting the development of the PROMs were included for the assessment of content validity.[100] Other studies were considered eligible for the assessment of content validity if they were full-text original articles, including adult patients (>18 years of age) with nonspecific LBP[67] and/or professionals (eg, researchers, clinicians) to assess the relevance, comprehensiveness, or comprehensibility of the content of ≥1 of the 3 PROMs.[100] Studies on all the other measurement properties were included if they were full-text articles presenting results for adult patients with nonspecific LBP. Studies in populations that also included patients with specific LBP or patients with pain locations different from the lower back were included only if ≥75% of the total sample was classified as having nonspecific LBP or if results were presented separately for the group with nonspecific LBP.[54] Studies that used the PROMs as outcome measurement instruments, or in which the PROMs were used in a validation studies of other instruments, were excluded.[84]

Inclusion criteria were applied by 2 reviewers (A.C. and L.M.) independently to the titles and abstracts of the hits retrieved with the searches. Potentially eligible full texts were screened independently by the same 2 reviewers. Consensus on inclusion was sought between reviewers and, in case of disagreement, a third reviewer (R.O.) made decisions.

Evaluation of the Measurement Properties

After retrieving the available evidence, COSMIN guidance for systematic reviews of PROMs recommends assessment of measurement properties in the following order: 1) content validity, 2) internal structure (ie, structural validity, internal consistency, and cross-cultural validity), and 3) the remaining properties (ie, test-retest reliability, measurement error, criterion validity, construct validity, responsiveness).[84] For each measurement property, 3 phases are included in the assessment. First, the risk of bias of each single study on a measurement property is assessed. Second, the results of each single study on a measurement property are rated against criteria for sufficient measurement properties. Third, the results from all studies on a measurement property are summarized and the quality of evidence is graded. Each phase is described in more detail in the following sections.

Risk of Bias Assessment and Data Extraction

The risk of bias of the included studies was assessed with the COSMIN Risk of Bias checklist.[72] Risk of bias refers to the methodologic quality of the studies. The COSMIN checklist contains a box for each measurement property and boxes to assess the PROM development quality.[100] Each box is rated on a 4-point rating scale: very good, adequate, doubtful, or inadequate. For the development study, total quality scores were determined separately for the 2 main parts of the study: concept elicitation study and cognitive interview(s) with patients. For the content validity studies, the study quality for the 3 main aspects of content validity (ie, relevance, comprehensiveness, comprehensibility) was assessed separately. A total rating was obtained for each part by taking the lowest rating among the standards (ie, worst score counts).[99] Two reviewers (A.C. and C.T.) assessed PROM development quality and the risk of bias of original content validity studies independently and achieved consensus in a face-to-face meeting.

A similar 4-point rating scale and worst score counts method were also used for assessing the risk of bias for studies on the other measurement properties[72] and a total quality rating was determined for the studies on each measurement property in each study. Two reviewers (A.C. and L.M.) assessed the risk independently and achieved consensus in a video conference. For every study, data was extracted on patient characteristics and results by 1 reviewer (A.C.) and checked for accuracy by a second reviewer (L.M.).

Evidence Synthesis

Evidence synthesis was performed separately for each measurement property.[84,100] For content validity, the results of the studies (including PROM development) were rated by 2 reviewers (A.C. and C.T.) independently according to 10 established criteria: 5 on relevance, 1 on comprehensiveness, and 4 on comprehensibility.[100] Each criterion could be rated as sufficient (+), insufficient (−), or indeterminate (?). The same criteria were also applied by 2 reviewers (A.C. and C.T.) to the content of the PROM itself[100]; a specific version of the VAS and NRS was used for this assessment, with the introductory question, recall period, and external descriptors as recommended in a recent consensus study (Appendix 2).[8] An overall sufficient (+), insufficient (−), or inconsistent (±) rating was determined for relevance, comprehensiveness, and comprehensibility of each PROM by jointly assessing all results and reviewers' ratings on the same PROM. More detailed information on this assessment can be found in the COSMIN user manual on assessing the content validity of PROMs (available at: www.cosmin.nl).

For the other measurement properties, the results were rated according to the consensus-based criteria proposed by Prinsen et al[85] (Appendix 3). For measurement error, consensus-based minimal important change values[75] were used to judge the relative magnitude of the smallest detectable change. For construct validity and responsiveness, the review team formulated a set of a priori hypotheses against which to evaluate the results of studies. For both properties, correlations were expected to be:

- ≥.60 with other pain intensity instruments;
- <.60 and ≥.30 with instruments measuring related but dissimilar constructs (eg, pain behavior, physical functioning); and
- <.30 with instruments measuring unrelated constructs.

These hypotheses were based on the results of a systematic review on physical functioning PROMs for LBP.[10] Two additional hypotheses were formulated for responsiveness:

- the area under the curve to discriminate between improved and not improved/deteriorated patients had to be ≥.70;
- effect sizes and standardized response means for improved patients had to be ≥.50 larger than those for not improved/deteriorated patients; the effect size referred to the mean difference divided by the baseline standard deviation, whereas the standardized response mean referred to mean differences divided by the standard deviation of the difference.[20]

For construct validity and responsiveness, an overall sufficient (+), insufficient (−), or inconsistent (±) rating was determined by counting the number of results that met the hypotheses across all studies.[84] For the other measurement properties, an overall rating was determined by lumping together the scoring of each individual study; if ≥75% of the studies displayed the same scoring, that scoring became the overall rating (+ or −), whereas if <75% of studies displayed the same scoring, the overall rating became inconsistent (±).[84]

The quality of evidence for each measurement property was rated according to the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) approach,[37] adapted for this type of review, into high, moderate, low, or very low.[84,100] High-quality evidence indicates that further research is very unlikely to change the confidence in study results; moderate indicates that is likely that further research will have an important impact on study results and may change them; low suggests that further research is very likely to have an important impact on study results and is likely to change them; very low means that any result is very uncertain.[37] For content validity, the evidence quality could be downgraded because of risk of bias and inconsistency of results and indirectness, as outlined elsewhere.[100] For the other measurement properties, risk of bias, imprecision, inconsistency, and indirectness were taken into account to rate the evidence quality.[84] The concepts of risk of bias, imprecision, inconsistency, and indirectness were taken from the GRADE approach.[37] Risk of bias refers to limitations in the methodologic quality of the eligible studies, imprecision refers to a low total number of patients included in the studies, inconsistency refers to unexplained heterogeneity of studies' results, and indirectness refers to the extent to which the study characteristics met the review inclusion criteria.[32]

Rating the quality of evidence for content validity was performed by giving more weight to original content validity studies over PROM development and reviewers' rating, as explained elsewhere (Appendix 4).[100] Thus, if there were no content validity and no PROM development studies (or if the PROM development was of inadequate quality), the overall rating corresponded to the reviewers' rating and quality of evidence was labelled as very low.[100] For the other measurement properties, downgrading was done for risk of bias of 1 level if there was only 1 adequate quality study, 2 levels if there were only doubtful or inadequate studies; imprecision of 1 level if the total patient sample was <100 and 2 levels if <50; inconsistency of 1 level if ≥75% of studies results were not all sufficient (+), insufficient (−), or inconsistent (±); indirectness of one level if ≥1 study did not specifically address the construct (pain intensity) or the target population (adult patients with nonspecific LBP) of this review (Appendix 4).[11]

Results

Among 10,719 records retrieved, 23 full-text articles were included, 5 of which retrieved through citation tracking (Fig. 1). Of 45 potentially eligible articles retrieved in the databases, 27 were excluded: 5 did not present results separately for patients with nonspecific LBP,[1,57,58,93,101] 9 did not aim to assess any measurement property,[18,27,28,34,35,40,48,49,92] 8 did not report clearly if patients with nonspecific LBP were included,[25,26,38,61,66,78,83,89] and one each was excluded for the following reasons: the VAS administered over the phone,[46] the VAS completed by a tester,[74] assessed patients with experimental pain,[82] assessed only patients with specific LBP,[87] and focus on other instruments.[109]

Three of the included full-text articles reported information on the BPI-PS development[15,16,19] and the other 20 included 22 original studies (2 articles included 2 studies each[36,59]) on the measurement properties of the 3 PROMs. The VAS was assessed in 10 studies, the NRS in 13, and the BPI-PS in 4. Four studies assessed >1 PROM for the same patient group[36,88,95] (Table 1).

VAS

A 100-mm VAS was used in all 10 studies; introductory statement, time recall period, and external verbal descriptors varied (Table 1). One study assessed content validity,[88] 2 test-retest reliability,[64,80] 2 measurement error,[76,80] 2 construct validity,[29,95] and 4 responsiveness.[3,36,91] Patients' characteristics of each study are presented in Table 1 and their results in Tables 2 to 4.

Content Validity

None of the studies retrieved described the development of the VAS as a pain intensity instrument. Robinson-Paap et al[88] assessed VAS relevance and comprehensiveness with adequate quality; the same

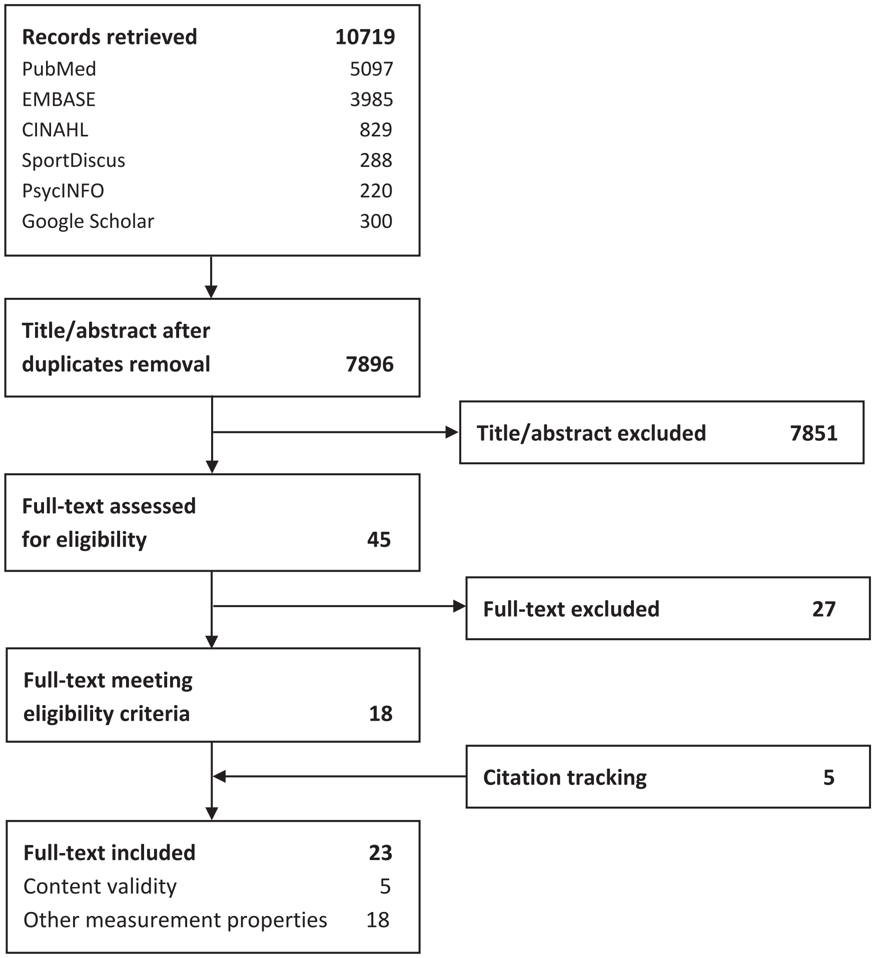

Figure 1. Flow chart of results of search strategy and selection of records.

study also assessed NRS and BPI-PS. Three main themes were identified by patients with LBP on the instruments: 1) perception that it may not even be possible to measure pain in a meaningful way, 2) difficulty in finding appropriate experiences as referents, and 3) difficulty with averaging pain. A few specifications for each theme are presented here.

1) Example: "At the end of the day a single line is really not going to tell what I' m actually feeling." Three more specific subthemes were identified:

a Pain measurement is influenced by other things other than pain.
b The numbers used to rate pain do not have an absolute meaning.
c Preference for pain intensity ratings in the middle of the scale.

2) This theme included 2 subthemes:

a Some patients used their prior LBP episodes as comparators; others did not use a comparator experience at all; rather, they thought of pain based on how much medication they took in a particular day.
b Several patients thought that anchoring the lower end to no pain was not appropriate because they always experience some pain. Some patients expressed that they would not use the highest numbers on the scale because doing so would indicate a lack of ability to cope with the pain. The suggestions of average, normal, or usual as alternative anchors also emerged.

3) Generating a number to represent average pain over a given time period was not an intuitive task. The longer the time period over which to average, the more difficulty participants had.

Relevance and comprehensiveness were rated as insufficient based on these results; the reviewers rated relevance, comprehensiveness, and comprehensibility of the VAS as sufficient. Low-quality evidence was found for inconsistent findings for relevance and comprehensiveness, owing to inconsistency and indirectness, because the only eligible study did not specifically focus on the pain intensity construct, but on pain in general without referring to a specific aspect such as intensity (Table 5). Very low-quality evidence was found for sufficient comprehensibility (Table 5).

Internal Structure

Structural validity and internal consistency are not applicable to the VAS and NRS because these measures are single-item instruments. No studies were found on cross-cultural validity.

Table 1. Characteristics of the Studies Included in This Systematic Review

PROM(s)	Reference	Language (Country)	Study Design	LBP Characteristics	Measurement Properties	PROM(s) Description	PROM Scores, $\mu \pm SD$	Pain Construct	High Anchor*	Patient Characteristics			
										N	Female, %	Age, Years, $\mu \pm SD$	Pain Duration, $\mu \pm SD$
VAS, NRS, BPI-PS	Robinson-Papp[88]	English (US)	Focus groups and individual interviews	>2 months with or without leg pain	Content validity	10-cm VAS 11-point NRS BPI-PS		Average past 24 h NA	Worst pain NA	13	54	45	
Two VASs, NRS	Strong[95]	English (Australia)	Cross sectional	Chronic	Construct validity	100-mm VAS 100-mm v-VAS 11-point NRS	60 ± 24 61 ± 24 6.3 ± 2.3	Intensity	Pain as bad as it could be	92	49	46 ± 13	10 ± 10 years
VAS, NRS	Grotle[36]	Norwegian	Longitudinal	<3 weeks	Responsiveness	100-mm VAS 11-point NRS	39 ± 23 6.8 ± 1.8	For the time being During the last week	Pain as bad as it could be	54	73	38 ± 10	10 ± 7 days
VAS, NRS	Grotle[36]	Norwegian	Longitudinal	>3 months	Responsiveness	100-mm VAS 11-point NRS	34 ± 23 6.1 ± 2.4	For the time being During the last week	Pain as bad as it could be	50	62	40 ± 9	2 ± 2 years
Three VASs	Love[64]	English (Australia)	Cross sectional	>6 months	Test–retest reliability	10-cm VAS 10-cm VAS 10-cm VAS		Experienced now At its worst At its best	Intolerable pain	63			
VAS	Beurskens[3]	Dutch	RCT	>6 weeks	Responsiveness	100-mm VAS		Average severity during last week		81	46	41 ± 10	24 weeks (median)
VAS	Ostelo[76]	Dutch	Cross sectional	<4 weeks with or without radiation (no pain ≥3 months before)	Measurement error	100-mm VAS		Current intensity	Worst imaginable pain	176	40	43 ± 12	1/3 each: <1 week, 1-2 weeks, 2-4 weeks
VAS	Sheldon[91]	English (US)	Two RCTs	>3 months with or without leg pain analgesic intake ≥24 d/mo	Responsiveness	100-mm VAS	77 ± 14	Intensity	Extreme pain	639	62	53 ± 13	11 ± 11 years

(continued on next page)

Table 1. (*Continued*)

PROM(s)	Reference	Language (Country)	Study Design	LBP Characteristics	Measurement Properties	PROM(s) Description	PROM Scores, $\mu \pm SD$	Pain Construct	High Anchor*	Patient Characteristics			
										N	Female, %	Age, Years, $\mu \pm SD$	Pain Duration, $\mu \pm SD$
VAS	Paungmali[80]	Thai	Cross sectional	>3 months VAS score = 2-7	Test–retest reliability, measurement error	10-cm VAS	39 ± 9	Average over the lumbosacral area	Extreme pain	13	69	26 ± 6	1 ± 1 years
VAS	Fishbain[29]	English (US)	Longitudinal	>6 months as primary complaint	Construct validity	100-mm v-VAS	62 ± 32	Current	Unbearable pain	236			
Four NRSs	Hush[44]	English (Australia)	Focus groups	Persistent or recurrent LBP, or recovery from previous LBP	Content validity	11-point NRS		At its worst in the past 24 h	Pain as bad as you can imagine	36	42	42 ± 6	69% persistent /recurrent, 31% recovery
								At its least in the past 24 h					
								On the average					
								Right now					
Three NRSs[†]	Childs[14]	English (US)	RCT	With or without leg symptoms, ODI ≥ 30%	Test–retest reliability, measurement error, responsiveness	11-point NRS	5.8 ± 2.0	Current level during last 24 h	Worst imaginable pain	131	42	34 ± 11	66% at <6 weeks
								Best level during last 24 h					
								Worst level during last 24 h					
NRS	Kovacs[56]	Spanish (Spain)	Longitudinal	>14 days, with or without leg pain NRS ≥ 3/10	Measurement error, responsiveness	11-point NRS	7.5 ± 2.0	Lower back	Worst imaginable pain	1349	68	54 ± 15	9 ± 8 years
NRS	Pengel[81]	English (Australia)	RCT	>6 weeks and <3 months	Responsiveness	11-point NRS	5.5 ± 2.1	Average over past week	Worst pain possible	156	56	49 ± 16	
NRS[‡]	Lauridsen[59]	Danish	Longitudinal	With or without leg pain	Responsiveness	11-point NRS	4.3 ± 2.3	Back pain with or without leg pain over past week	Worst possible pain	94	53	44	73% ≤30 days, rest >30 days

(*continued on next page*)

Table 1. (*Continued*)

PROM(s)	Reference	Language (Country)	Study Design	LBP Characteristics	Measurement Properties	PROM(s) Description	PROM Scores, $\mu \pm SD$	Pain Construct	High Anchor*	Patient Characteristics			
										N	Female, %	Age, Years, $\mu \pm SD$	Pain Duration, $\mu \pm SD$
NRS[§]	Lauridsen[59]	Danish	Longitudinal	With or without leg pain	Responsiveness	11-point NRS	4.9 ± 2.5	Back ± leg pain over past week	Worst possible	97	54	47	12% ≤ 30 days, rest 30 days
NRS[¶]	Van der Roer[104]	Dutch	RCT		Measurement error	11-point NRS	6.4 ± 1.8	Intensity	Very severe pain		114		
NRS	Lauridsen[60]	Danish	Longitudinal	With or without leg pain	Measurement error	11-point NRS	6.2	Intensity over past week	Worst possible pain	147	66	46	37% at ≤ 6 months
NRS	Maughan[69]	English (UK)	Longitudinal	>3 months with or without leg pain	Test–retest reliability, measurement error, responsiveness	11-point NRS	5.0 ± 2.6	Intensity	Worst imaginable pain	48	67	52	6 years (mean)
BPI-PS	Keller[55]	English (US)	Longitudinal		Internal consistency, construct validity, responsiveness	BPI-PS		NA	NA	131	50	46 ± 14	
BPI-PS	Tan[97]	English (US)	Cross-sectional	Chronic	Internal consistency, Structural validity, Construct validity	BPI-PS	7.0 ± 1.8	NA	NA	440	8	55	10 ± 7 days
BPI-PS	Whynes[106]	English (UK)	RCT		Responsiveness	BPI-PS	8.1 ± 3.0	NA	NA	37			

Abbreviations: SD, standard deviation; v-VAS, vertical VAS; RCT, randomized controlled trial; ODI, Oswestry Disability Index; NA, not applicable.
Note. Empty cells reflect data not assessed.
* The low anchor was always no pain.
† The average of the 3 ratings was used to represent the patient's overall pain intensity.
‡ This study refers to primary care patients.
§ This study refers to secondary care patients.
¶ Measurement error was calculated on unchanged patients but characteristics of those patients alone were not presented.
‖ These are scores were the same for patients with (sub)acute LBP or chronic LBP.

Table 2. Test-Retest Reliability and Measurement Error of Pain Intensity Instruments in Patients With LBP

PROM(s)	Reference	Pain construct	Test-Retest Reliability n	Study Quality	Time Interval(s)	ICC (95% CI)	Measurement Error n	Study Quality	Time Interval(s)	SEM (95% CI, % Scale Range)	SDC* (95% CI, % Scale Range)
Three VASs	Love[64]	Experienced now	63	Doubtful	Some days	.77†					
		At its worst				.49†					
		At its best				.57†					
VAS	Ostelo[76]	Current intensity					176	Doubtful	Maximum 24 hours	13 (12-15, 13)‡	36 (32-41, 36)‡
VAS	Paungmali[80]	Average over the lumbosacral area	13	Doubtful	48 hours	.90‡	13	Inadequate	48 hours	.1 (—, 1)‡	.3 (—, 3)§
Three NRSs*	Childs[14]	Current, best, and worst level during last 24 h	41	Adequate	1 week	.61 (.30-.77)‡	41	Adequate	1 week	1.0 (—, 10)‡	2.8 (—, 28)§
NRS	Kovacs[56]	Lower back					209¶	Adequate	12 weeks	1.3 (—, 13)§	3.5 (3.2-3.8, 35)
NRS	van der Roer[104]	Intensity					52$^{\|}$	Doubtful	12 weeks	1.7 (—, 17)§	4.7 (3.3-8.0, 47)
							62$^{\|}$			1.6 (—, 16)§	4.5 (3.4-6.7, 45)
NRS	Lauridsen[60]	Intensity over past week					55	Adequate	1 week	1.0 (—, 10)§	2.8 (—, 28)
NRS	Maughan[69]	Intensity	25	Adequate	5 weeks	.92†	25	Adequate	5 weeks	.9 (—, 9)‡	2.4 (—, 24)‡

Abbreviations: ICC, intraclass correlation coefficient; SEM, standard error of measurement; SDC, smallest detectable change.
Note. Empty cells represent aspects not assessed.
* The average of the 3 ratings was used to represent the patient's overall pain intensity.
† This value represents a Pearson product-moment correlation and not an intraclass correlation coefficient.
‡ It is unclear if $ICC_{consistency}$ $SEM_{consistency}$, or $ICC_{agreement}$, $SEM_{agreement}$ was used.
§ This SEM or SDC was not reported in the article but it was calculated from the available data (SDC was calculated as SEM x $\sqrt{2} \times 1.96$).
¶ The sample size for the measurement error of the NRS for LBP was not reported in the article; therefore, this number includes also patients with leg pain.
‖ There were 52 patients with (sub)acute LBP, and 62 patients with chronic LBP.

Table 3. Construct Validity (Hypotheses Testing) of Pain Intensity Instruments in Patients With LBP

PROM(s)	Reference	Pain Construct	N	Study Quality	Correlations With Other Measurement Instruments Measuring Similar, Related, or Unrelated Constructs
Two VAS, NRS	Strong[95]	Intensity	92	Inadequate	see below
VAS	Fishbain[29]	Current lower back	236	Adequate	.17 with pain thresholds; .29 with pain tolerance
BPI-PS	Keller[55]	Intensity	131	Adequate	see below
BPI-PS	Tan[97]	Intensity	440	Very good	.40 with Roland Morris Disability Questionnaire

Strong[95]:

	NRS	VAS	v-VAS	BRS	VRS	NRS-101	PPI	PRI
NRS		.81	.70	.53	.71	.85	.51	.20
VAS	.81		.81	.50	.64	.81	.48	.22
v-VAS	.70	.71		.43	.54	.73	.45	.25

Keller[55]:

CPG IS	CPG DS	RMDQ	SF-36 BP	SF-36 PF	SF-36 RP	SF-36 GH	SF-36 V	SF-36 SF	SF-36 RE	SF-36 MH
.60	.49	.57	.61	.63	.54	.37	.47	.51	.41	.41

Abbreviations: VAS, horizontal VAS; v-VAS, vertical VAS; BRS, Behavioral Rating Scale; VRS, Verbal Rating Scale with 4 response options (eg, no pain, some pain); NRS-101, NRS in which the patient should choose a number between 0 and 100 that indicates his or her level of pain; PPI, Present Pain Intensity ranging from 1 (mild) to 5 (excruciating); PRI, Pain Rating Index of the McGill Pain Questionnaire; MDQ, Roland Morris Disability Questionnaire; CPG-IS, Intensity Scale of the Chronic Pain Grade; CPG-DS, Disability Scale of the Chronic Pain Grade; SF36-BP, Bodily Pain subscale of the Short Form 36; SF36-PF, Physical functioning subscale of the Short Form 36; SF36-RP, Role Physical subscale of the Short Form 36; SF36-GH, General Health subscale of the Short Form 36; SF36-V, Vitality subscale for the Short Form 36; SF36-SF, Social Functioning subscale of the Short Form 36; SF36-RE, Role Emotional subscale of the Short Form 36; SF36-MH, Mental Health subscale of the Short Form 36.

Other Measurement Properties

Only 1 study[80] presented results that could be rated for test-retest reliability (Table 2), providing low-quality evidence (owing to risk of bias and imprecision) of sufficient reliability (Table 5). Owing to risk of bias and inconsistency of results across studies (Table 2), very-low quality evidence of inconsistent findings was found for measurement error (Table 5).

Results on hypothesis testing for construct validity were inconsistent across studies (Table 3), providing low-quality evidence (owing to risk of bias and inconsistency) on this measurement property (Table 5). The results of 4 studies were tested against our hypotheses for responsiveness (Table 4), providing low-quality evidence (owing to risk of bias and inconsistency of results) of inconsistent results for this measurement property (Table 5).

NRS

The 11-point NRS was used in all 13 studies; external descriptors varied slightly, whereas construct and recall period in the introductory statement varied more widely (Table 1). One study[14] administered 3 NRSs referring to current, best, and worst pain over the last 24 hours and took the average of the 3 scores in the analyses. Two studies evaluated content validity,[44,88] 2 test-retest reliability,[14,69] 5 measurement error,[14,56,60,69,104] 1 construct validity,[95] and 8 responsiveness[14,36,56,59,69,81]; 4 studies assessed the NRS in conjunction with other pain intensity instruments.[36,88,95]

Content Validity

No studies presenting the NRS development were found. Robinson-Paap et al[88] analyzed the NRS together with the VAS and BPI-PS, displaying the same results for all the instruments, as summarized for the VAS results. Hush et al[44] assessed the relevance and comprehensiveness of 4 NRS versions in a study of adequate quality. The majority of patients included in this study (ie, >50%) expressed the opinion that the NRS does not adequately capture the complexity of their personal experience of pain. Two themes emerged: 1) the meaning attributed to the pain score and 2) the time-frame of measurement. Regarding the first theme, participants reported that their score reflects many aspects of the pain experience, other than the sensory component of pain; another common view was that NRS scores are highly dependent on individual experiences of pain that can determine the benchmark used by a patient to rate the pain. Regarding the second theme, a majority believed that the NRS versions assessing pain in the past 24 hours or right now were unlikely to capture improvements because of symptom fluctuation.

These results, taken together with the reviewers' ratings on the NRS to measure pain intensity in LBP, provided inconsistent results based on low quality evidence (owing to inconsistency and indirectness; Table 5).

Table 4. Responsiveness (Hypotheses Testing) of Pain Intensity Instruments in Patients With LBP

PROM(s)	Ref	Study Quality	Time Interval	Criterion	PROM	Pain Construct	N	Better, Same, Worse (%)	Correlation With Criterion	AUC % (95% CI)	ESs* or SRMs† (95% CI)	Correlations With Changes In Other Instruments
VAS, NRS	Grotle[36]	Doubtful	4 weeks	6-point GPES from worse to completely recovered	VAS	For the time being	42	74 better, 26 same	.59	91 (83- 100)	.7 (.4 to 1.0) SRM overall; 1.6 (1.1 to 2.0) SRM better; -.5 (-.8 to .5) SRM same	.64 to RMDQ; .59 to ODI; .49 to DRI; .67 to SF36-PF; .65 to NRS
					NRS	During last week	45	76 better 24 same	.76	93 (86 to 100)	1.1 (.8 to 1.5) SRM overall; 2.0 (1.4 to 2.6) SRM better; 1.0 (.6 to 1.7) SRM same	.68 to RMDQ; .58 to ODI; .58 to DRI; .38 to SF36-PF; .65 to VAS
VAS, NRS	Grotle[36]	Doubtful	3 months	6-point GPES from worse to completely recovered	VAS	For the time being	33	48 better, 52 same	.24	71 (54 to 88)	-.1 (.4 to 1.0) SRM overall; .4 (-.2 to .9) SRM better; .1 (-1.1 to .3) SRM same	.40 to RMDQ; .35 to ODI; .13 to DRI; -.08 to SF36-PF; .30 to NRS
					NRS	During last week	39	49 better, 51 same	.52	82 (67 to 96)	.3 (.0 to .6) SRM overall; 1.1 (.4 to 1.7) SRM better; -.2 (-.6 to .4) SRM same	.52 to RMDQ; .42 to ODI; .16 to DRI; .13 to SF36-PF; .30 to VAS
VAS	Beurskens[3]	Adequate	5 weeks	7-point GPES from completely recovered to vastly worsened	VAS	Average severity during last week	81‡	47 better, 48 same, 6 worse		91	1.6 SRM better; .1 SRM same	
VAS	Sheldon[91]	Doubtful	12 weeks	5-point PGART from excellent to none	VAS	Lower back intensity	639		.68-.74§	88 (85 to 90)	1.8-2.6 ES overall§	.66-.70 to RMDQ§
NRS	Pengel[81]	Doubtful	6 weeks	11-point GPES from vastly worse to completely recovered	NRS	Average over past week	156		.50		1.3 (1.2 to 1.4) ES overall‡	
Three NRSs¶	Childs[14]	Doubtful	1 week	15-point RS from a great deal worse to a very great deal better‖	NRS	Current, best, and worst level during last 24 h	131††	65 better, 33 same, 2 worse		72 (62 to 81)	.9 SRM overall; 1.4 SRM better; .5 SRM same	
			4 weeks					82 better, 13 same, 4 worse		92 (86 to 97)	1.2 SRM overall; 1.5 SRM better; .6 SRM same	
NRS	Lauridsen[59]	Adequate	8 weeks	7-point GPES from much better to much worse, and NRS to score pain change importance	NRS	Back and/or leg over past week	85#	73 better, 27 same		65 in LBP only	1.5 (1.2 to 1.8) SRM better; .8 (.3 to 1.3) SRM same	
NRS	Lauridsen[59]	Adequate	8 weeks		NRS		59**			62 in LBP only		

(continued on next page)

Table 4. (*Continued*)

PROM(s)	Ref	Study Quality	Time Interval	Criterion	PROM	Pain Construct	n	Better, Same, Worse (%)	Correlation With Criterion	AUC % (95% CI)	ESs* or SRMs† (95% CI)	Correlations With Changes In Other Instruments
				7-point GPES from much better to much worse, and NRS to score pain change importance		Back and/or leg over past week		31 better, 69 same			.9 (.4 to 1.3) SRM better; .2 (-.1 to .5) SRM same	
NRS	Kovacs[56]	Doubtful	12 weeks	4-point RS from completely recovered to worsened	NRS	Lower back	1349	33 recovered 50 better 16 same 1 worse		95 (93 to 97)	3.2 SRM recovered‖; -2.0 SRM improved‖; -.5 SRM unchanged‖; 1.6 SRM deteriorated‖	
NRS	Maughan[69]	Doubtful	5 weeks	7-poing GPES from completely recovered to vastly worsened	NRS	Intensity	48	48 better 52 same		50		
BPI-PS	Keller[55]	Inadequate		RMDQ	BPI-PS	NA	131	34 better 50 same 16 worse			-1.1 SRM improved; -.4 SRM unchanged; .3 SRM deteriorated	
BPI-PS	Whynes[106]	Inadequate	12 weeks		BPI-PS	NA	37				.9 (.8-1.0) SRM overall	.66 with BPI-PI; .70 with ODI; -.57 with EQ5D-US; -.56 with EQ5D-VAS

Abbreviations: AUC, area under the ROC curve; SRM, standardized response mean; GPES, global perceived effect scale; RMDQ, Roland Morris Disability Questionnaire; ODI, Oswestry Disability Index; DRI, Disability Rating Index; SF36-PF, physical functioning subscale of the Short Form 36; PGART, patient global assessment of response to therapy; RS, rating scale; BPI-PI, pain interference subscale of the BPI; NA , not applicable; EQ5D-US, utility score of the EuroQol-5D; EQ5D-VAS, VAS of the EuroQol-5D.

Note. Empty cells indicate not available or not assessed data.

* ESs were calculated by dividing the mean change by the baseline standard deviation.

† SRMs were calculated by dividing the mean change by its standard deviation.

‡ In this case, an 84% CI was presented.

§ This is the range of correlations or ESs found in the 3 separate arms of this study (ie, etoricoxib 60 mg, etoricoxib 90 mg, placebo).

¶ The average of the 3 ratings was used to represent the patient's overall pain intensity.

‖ These ESs or SRMs were not reported in the article but calculated from the available data.

\# Primary care patients.

** Secondary care patients.

†† There were 125 patients who completed the 1-week follow-up and 119 patients the 4-week follow-up.

Table 5. Evidence Synthesis on Measurement Properties of Pain Intensity Instruments in Patients with LBP

Measurement Properties			VAS	NRS	BPI-PS
Content validity	Relevance	Rating	±	±	±
		Quality of evidence	Low	Low	Low
	Comprehensiveness	Rating	±	±	±
		Quality of evidence	Low	Low	Low
	Comprehensibility	Rating	+	+	+
		Quality of evidence	Very low	Very low	Very low
Structural validity		Rating	NA	NA	+
		Quality of evidence			Moderate
Internal consistency		Rating	NA	NA	+
		Quality of evidence			Moderate
Test-retest reliability		Rating	+	±	
		Quality of evidence	Very Low	Low	
Measurement error		Rating	±	–	
		Quality of evidence	Very Low	High	
Construct validity		Rating	±	±	±
		Quality of evidence	Low	Very Low	Moderate
Responsiveness		Rating	±	±	±
		Quality of evidence	Low	Moderate	Very Low

Abbreviations: +, sufficient results; –, insufficient results; ±, inconsistent results; NA, measurement property not applicable.
Note. Empty cells represent measurement properties not assessed in any study. The cross-cultural validity row is not displayed because it was not assessed in any study.

Internal Structure

Structural validity and internal consistency are not applicable to the NRS because it is a single-item scale and no studies assessing cross-cultural validity were retrieved.

Other Measurement Properties

Low-quality evidence (owing to inconsistency and imprecision) was found for inconsistent findings for test-retest reliability (Tables 2 and 5). High-quality evidence was found for insufficient measurement error (Table 5) because the smallest detectable change values in 4 adequate quality studies were greater than the proposed 2-point minimal important change (Table 2).[75]

Very low-quality evidence from 1 study of inadequate quality was found for inconsistent results on construct validity (Tables 3 and 5). Seven of the 8 responsiveness studies provided results to be rated against our hypotheses (Table 4), resulting in inconsistent results based on moderate quality evidence (owing to inconsistency; Table 5).

BPI-PS

Three studies presented information on the BPI-PS development.[15,16,19] Among the other 4 studies (Table 1), 1 assessed content validity,[88] 2 internal consistency,[55,97] 1 structural validity,[97] 2 construct validity[55,97] and 2 responsiveness.[55,106]

Content Validity

The development of the BPI was rated as of doubtful quality because it was unclear if the included patients were representative of the target population.[11] One content validity study assessed relevance and comprehensiveness in a study of adequate quality.[88] This study also assessed the VAS and the NRS, providing the same results for all 3 instruments, as outlined elsewhere in this article. It was considered to provide indirect evidence because the pain intensity construct was not clearly specified and its negative results were in contrast with reviewers' ratings; this resulted in low-quality evidence for inconsistent findings (Table 5).

Internal Structure

One study[97] assessed the BPI-PS structural validity in a study of adequate quality performing an exploratory factor analysis on the whole BPI. The 4 BPI-PS items loaded on the same factor explaining 12% of the total variance and with eigenvalue equal to 1.38. The factor loadings on this factor ranged from .61 (pain worst) to .82 (pain least), whereas factor loadings on the first pain interference factor were very low (between -.07 and .16). This finding resulted in sufficient unidimensionality based on moderate quality evidence (Table 5).

Two studies of adequate quality investigated the internal consistency, exhibiting Cronbach's alpha values of .82[55] and .85.[97] According to the latest COSMIN guidance,[84] these results provide moderate quality evidence for sufficient internal consistency (Table 5). No studies on cross-cultural validity were retrieved.

Other Measurement Properties

Test-retest reliability and measurement error of the BPI-PS were not assessed in any study. Moderate quality evidence (owing to inconsistent results across studies; Table 3) was found for inconsistent results on construct validity (Table 5). Responsiveness was assessed in 2 studies of inadequate quality (Table 4), providing very low-quality evidence (owing to risk of bias and inconsistency) of inconsistent results for this measurement property (Table 5).

Discussion

This systematic review illustrates that the quality of evidence on the measurement properties of the VAS, NRS, and BPI-PS in patients with LBP is clearly suboptimal (Table 5). The quality of evidence on content validity of all 3 instruments is low to very low. For the other measurement properties, high-quality evidence was only found on the insufficient measurement error of the NRS. Moderate quality evidence was found for inconsistent results on the NRS responsiveness, sufficient results for BPI-PS structural validity and internal consistency, and inconsistent construct validity of the BPI-PS. For all other assessed measurement properties, the quality of evidence was low or very low (Table 5).

The NRS is most often recommended to measure pain intensity in patients with LBP[13,17,22] and in chronic pain more generally.[24] Apparently, only practical aspects have dictated NRS recommendations in LBP so far. In a recent international Delphi survey, researchers, clinicians, and patients clearly preferred the NRS over VAS and BPI-PS to measure pain intensity in LBP clinical trials.[8] Several Delphi participants highlighted the VAS to be less understandable for patients (the elderly in particular) than the NRS, time consuming to score if the line is not exactly 100 mm long, and difficult to administer with digital devices.[8] Meanwhile, the BPI-PS was less often chosen because it has a fee for administration and it is less easy to administer than the other instruments.[8] A previous review on a broader pain population also concluded that the NRS was preferred over the VAS for feasibility reasons.[42] Despite these preference toward the NRS, the VAS has been the most frequently used pain instrument in LBP clinical trials so far[31]; therefore, it is important to monitor if this pattern of use will change in the (near) future.

Content validity is considered the first measurement property to consider when selecting a PROM.[85] Evidence on this property could be generated by head-to-head comparison studies where all 3 instruments are administered and patients are asked to rate their relevance, comprehensiveness, and comprehensibility.[100] Two studies included in this review[44,88] raised issues regarding the content validity of NRS and VAS, in line with the results of a previous study in a chronic pain population.[107] If these results are replicated in future studies in patients with LBP, the use of these instruments should be seriously reconsidered. Because these PROMs are usually intended to measure pain intensity, future clinimetric studies should consider NRS and VAS versions that specifically refer to pain intensity in the introductory question, as displayed in Appendix 2. Structural validity and internal consistency of the BPI-PS were found to be sufficient (Table 5), which is not surprising considering that the BPI-PS items share very similar content; this could artificially inflate its unidimensionality and Cronbach's alpha.

This systematic review clearly showed that the NRS measurement error is larger than the 2-point minimal important change value commonly proposed for this instrument in LBP (Table 3).[75] This finding implies that this PROM may not be able to distinguish the smallest detectable changes from real changes in the measured construct,[21] which represents a serious limitation. Whether or not VAS and BPI-PS share this problem is not able to be determined because direct comparisons are lacking. The measurement error of an instrument can be decreased by increasing the number of repeated measurements or items,[20] as recently shown in mixed chronic pain populations—multi-item tools displayed slightly more reliable scores than single-item tools[50,51]; therefore, the BPI-PS may also have a smaller measurement error than the other 2 PROMs in patients with LBP, but this has to be tested.

The cross-cultural validity of the VAS, NRS, and BPI-PS has not been evaluated in patients with LBP or in broader populations with pain. Because data for patients with LBP from different cultures are routinely pooled in systematic reviews of clinical trials[54,65,68,79] and observational studies,[30,62] it is essential to exclude substantial differential item functioning across countries and languages. The evidence quality on construct validity and responsiveness is low (Table 5) to determine if any instrument outperforms the others. The only study directly comparing construct validity of VAS and NRS is of inadequate quality.[95] Two studies (of doubtful quality) comparing VAS and NRS responsiveness showed that the NRS has larger effect sizes (and, therefore, a better ability to capture pain intensity changes) in patients with acute and chronic LBP,[36] but this finding requires replication. There is evidence that multiple-item PROMs for pain do not display substantially larger effect sizes than single-item ones in more heterogeneous pain conditions,[48,50] but these studies did not specifically include the BPI-PS and did not specifically assess a range of responsiveness aspects, such as the area under the curve and correlations with other instruments.

Recently, the use of pain intensity scales in patients with chronic pain has been criticized.[2,63,96] More specifically, these instruments have been advocated as potential contributors to the opioid epidemic in some countries; patients who display high pain intensity ratings are those who, despite the presence of comorbidities such as mental health disorders, are frequently prescribed opioids, resulting in subsequent addiction.[2,96] Additionally, it has been proposed that "zero pain is not the (only) goal" in patients with chronic pain; rather, the main goal should be to improve (physical and psychological) functioning.[2,63] This view against the use of pain intensity scales and on the unimportance of pain intensity is in contrast with various studies clearly showing that decreasing pain intensity is a crucial goal for patients living with chronic pain.[9,41,43,70] Therefore, considering the importance of pain intensity as a core outcome domain in LBP[9] and considering that the instruments included in this review have been widely used for decades,[7,31] the lack of robust evidence supporting the measurement properties of the most frequently used instruments for this domain is worrisome. Nevertheless, it should be underlined that the GRADE approach in systematic reviews on measurement properties of instruments has only recently been introduced[85,100] and this is the first systematic review to adopt such an approach for all measurement properties; therefore, reaching the high-quality evidence level will be the goal of future research.

There is a need for adequate quality head-to-head comparison studies on pain intensity instruments in patients with LBP. The instruments assessed in this review may be included in these studies alongside other pain intensity instruments, such as Verbal Rating Scales, the bodily pain subscale of the Short Form 36 (which combines pain intensity measurement with pain interference), or other pain items or subscales of other generic- or disease-specific instruments. Additionally, other methods to assess pain intensity in patients with pain may be considered and investigators with innovative and creative ideas on how to better measure pain intensity are certainly welcome in this field.

The main strength of this first systematic review on the measurement properties of the 3 most frequently used pain intensity PROMs in LBP[7,31] is the use of the most up-to-date methodology.[72,73,84,100] In contrast with previous reviews on the measurement properties of pain intensity instruments,[6,42,47,52,86,94,108] this systematic review focused on patients with LBP only; this decision was guided by the focus of the core outcome measurement set for which this review was performed[8,9] and by the fact that there is evidence clearly suggesting what is the best method to synthesize the evidence on measurement properties of instruments (ie, whether it should be synthesized in specific or generic populations). A potential limitation is that the evidence synthesis lumps together studies from different languages and countries and includes instruments with (slightly) different pain constructs and high external anchors. However, this approach is routine for pain intensity scales in systematic reviews for LBP, splitting studies may be equally contentious, and there is no evidence on the best approach. For detailed scrutiny, language, country and instruments' characteristics of each study are specified in the results (Tables 1 to 4).

In conclusion, there is currently no evidence to claim superior measurement properties for any of the 3 commonly used instruments to measure pain in LBP. In our opinion, such evidence should preferably come from sound head-to-head comparison clinimetric studies, with priority to be given to the assessment of content validity, test-retest reliability, measurement error, and responsiveness.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jpain.2018.07.009.

References

1. Angst F, Verra ML, Lehmann S, Aeschlimann A: Responsiveness of five condition-specific and generic outcome assessment instruments for chronic pain. BMC Med Res Methodol 8:26, 2008

2. Ballantyne JC, Sullivan MD: Intensity of chronic pain: The wrong metric? N Engl J Med 373:2098-2099, 2015

3. Beurskens AJ, de Vet HC, Koke AJ: Responsiveness of functional status in low back pain: A comparison of different instruments. Pain 65:71-76, 1996

4. Black N: Patient reported outcome measures could help transform healthcare. BMJ 346:f167, 2013

5. Boers M, Brooks P, Strand CV, Tugwell P: The OMERACT filter for outcome measures in rheumatology. J Rheumatol 25:198-199, 1998

6. Castarlenas E, Jensen MP, von Baeyer CL, Miro J: Psychometric properties of the numerical rating scale to assess self-reported pain intensity in children and adolescents: A systematic review. Clin J Pain 33:376-383, 2017

7. Chapman JR, Norvell DC, Hermsmeyer JT, Bransford RJ, DeVine J, McGirt MJ, Lee MJ: Evaluating common outcomes for measuring treatment success for chronic low back pain. Spine 36:S54-S68, 2011

8. Chiarotto A, Boers M, Deyo RA, Buchbinder R, Corbin TP, Costa LO, Foster NE, Grotle M, Koes BW, Kovacs FM, Lin CC, Maher CG, Pearson AM, Peul WC, Schoene ML, Turk DC, van Tulder MW, Terwee CB, Ostelo RW: Core outcome measurement instruments for clinical trials in non-specific low back pain. Pain 159:481-495, 2018

9. Chiarotto A, Deyo RA, Terwee CB, Boers M, Buchbinder R, Corbin TP, Costa LO, Foster NE, Grotle M, Koes BW, Kovacs FM, Lin CC, Maher CG, Pearson AM, Peul WC, Schoene ML, Turk DC, van Tulder MW, Ostelo RW: Core outcome domains for clinical trials in non-specific low back pain. Eur Spine J 24:1127-1142, 2015

10. Chiarotto A, Maxwell LJ, Terwee CB, Wells GA, Tugwell P, Ostelo RW: Roland-Morris Disability Questionnaire and Oswestry Disability Index: Which has better measurement properties for measuring physical functioning in nonspecific low back pain? Systematic review and meta-analysis. Phys Ther 96:1620-1637, 2016

11. Chiarotto A, Ostelo RW, Boers M, Terwee CB: A systematic review highlights the need to investigate the content validity of patient-reported outcome measures for physical functioning in low back pain. J Clin Epidemiol 95:73-93, 2018

12. Chiarotto A, Terwee CB, Deyo RA, Boers M, Lin C-WC, Buchbinder R, Corbin TP, Costa LO, Foster NE, Grotle M, Koes BW, Kovacs FM, Maher CG, Pearson AM, Peul WC, Schoene ML, Turk DC, van Tulder MW, Ostelo RW: A core outcome set for clinical trials on non-specific low back pain: Study protocol for the development of a core domain set. Trials 15:511, 2014

13. Chiarotto A, Terwee CB, Ostelo RW: Choosing the right outcome measurement instruments for patients with low

back pain. Best Pract Res Clin Rheumatol 30:1003-1020, 2016

14. Childs JD, Piva SR, Fritz JM: Responsiveness of the numeric pain rating scale in patients with low back pain. Spine 30:1331-1334, 2005

15. Cleeland CS: The Brief Pain Inventory, Available at: https://www.mdanderson.org/documents/Departments-and-Divisions/Symptom-Research/BPI_UserGuide.pdf, 2009. Accessed July 27, 2017

16. Cleeland CS, Ryan K: Pain assessment: Global used of the Brief Pain Inventory. Ann Acad Med Singapore 23:129-138, 1994

17. Clement RC, Welander A, Stowell C, Cha TD, Chen JL, Davies M, Fairbank JC, Foley KT, Gehrchen M, Hagg O, Jacobs WC, Kahler R, Khan SN, Lieberman IH, Morisson B, Ohnmeiss DD, Peul WC, Shonnard NH, Smuck MW, Solberg TK, Stromqvist BH, Hooff ML, Wasan AD, Willems PC, Yeo W, Fritzell P: A proposed set of metrics for standardized outcome reporting in the management of low back pain. Acta Orthop 86:523-533, 2015

18. Co YY, Eaton S, Maxwell MW: The relationship between the St. Thomas and Oswestry disability scores and the severity of low back pain. J Manipulative Physiol Ther 16:14-18, 1993

19. Daut RL, Cleeland CS, Flanery RC: Development of the Wisconsin Brief Pain Questionnaire to assess pain in cancer and other diseases. Pain 17:197-210, 1983

20. de Vet HC, Terwee CB, Mokkink LB, Knol DL: Measurement in medicine: A practical guide. Cambridge, Cambridge University Press, 2011

21. de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM: Minimal changes in health status questionnaires: Distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes 4:54, 2006

22. Deyo RA, Dworkin SF, Amtmann D, Andersson G, Borenstein D, Carragee E, Carrino J, Chou R, Cook K, DeLitto A, Goertz C, Khalsa P, Loeser J, Mackey S, Panagis J, Rainville J, Tosteson T, Turk D, Von Korff M, Weiner DK: Report of the NIH Task Force on research standards for chronic low back pain. J Pain 15:569-585, 2014

23. Downie W, Leatham P, Rhind V, Wright V, Branco J, Anderson J: Studies with pain rating scales. Ann Rheum Dis 37:378-381, 1978

24. Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, Kerns RD, Stucki G, Allen RR, Bellamy N, Carr DB, Chandler J, Cowan P, Dionne R, Galer BS, Hertz S, Jadad AR, Kramer LD, Manning DC, Martin S, McCormick CG, McDermott MP, McGrath P, Quessy S, Rappaport BA, Robbins W, Robinson JP, Rothman M, Royal MA, Simon L, Stauffer JW, Stein W, Tollett J, Wernicke J, Witter J: Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 113:9-19, 2005

25. Elfving B, Lund I, C LB, Bostrom C: Ratings of pain and activity limitation on the visual analogue scale and global impression of change in multimodal rehabilitation of back pain - analyses at group and individual level. Disabil Rehabil 38:2206-2216, 2016

26. Farrar JT, Young Jr JP, LaMoreaux L, Werth JL, Poole RM: Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain 94:149-158, 2001

27. Filho IT, Simmonds MJ, Protas EJ, Jones S: Back pain, physical function, and estimates of aerobic capacity: What are the relationships among methods and measures? Am J Phys Med Rehabil 81:913-920, 2002

28. Fishbain DA, Gao J, Lewis JE, Zhang L: At completion of a multidisciplinary treatment program, are psychophysical variables associated with a VAS improvement of 30% or more, a minimal clinically important difference, or an absolute VAS score improvement of 1.5 cm or more? Pain Med 17:781-789, 2016

29. Fishbain DA, Lewis JE, Gao J: Is there significant correlation between self-reported low back pain visual analogue scores and low back pain scores determined by pressure pain induction matching? Pain Pract 13:358-363, 2013

30. Fritsch CG, Ferreira ML, Maher CG, Herbert RD, Pinto RZ, Koes B, Ferreira PH: The clinical course of pain and disability following surgery for spinal stenosis: A systematic review and meta-analysis of cohort studies. Eur Spine J 26:324-335, 2017

31. Froud R, Patel S, Rajendran D, Bright P, Bjorkli T, Buchbinder R, Eldridge S, Underwood M: A systematic review of outcome measures use, analytical approaches, reporting methods, and publication volume by year in low back pain trials published between 1980 and 2012: Respice, adspice, et prospice. PLoS One 11, 2016:e0164573

32. Furlan AD, Malmivaara A, Chou R, Maher CG, Deyo RA, Schoene M, Bronfort G, Van Tulder MW: 2015 updated method guideline for systematic reviews in the Cochrane Back and Neck Group. Spine 40:1660-1673, 2015

33. GBD 2015 Disease and Injury Incidence and Prevalence Collaborators: Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: A systematic analysis for the Global Burden of Disease Study 2015. Lancet 388:1545-1602, 2016

34. Gronblad M, Hurri H, Kouri JP: Relationships between spinal mobility, physical performance tests, pain intensity and disability assessments in chronic low back pain patients. Scand J Rehabil Med 29:17-24, 1997

35. Gronblad M, Lukinmaa A, Konttinen YT: Chronic low-back pain: Intercorrelation of repeated measures for pain and disability. Scand J Rehabil Med 22:73-77, 1990

36. Grotle M, Brox JI, Vollestad NK: Concurrent comparison of responsiveness in pain and functional status measurements used for patients with low back pain. Spine 29:E492-E501, 2004

37. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schunemann HJ: GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. BMJ 336:924, 2008

38. Hagino C, Thompson M, Advent J, Rivet L: Agreement between 2 pain visual analogue scales, by age and area of complaint in neck and low back pain subjects: The standard pen and paper VAS versus plastic mechanical sliderule VAS. J Can Chiropr Assoc 40:220, 1996

39. Hawker GA, Mian S, Kendzerska T, French M: Measures of adult pain: Visual Analog Scale for Pain (VAS Pain), Numeric Rating Scale for Pain (NRS Pain), McGill Pain Questionnaire (MPQ), Short-Form McGill Pain Questionnaire

(SF-MPQ), Chronic Pain Grade Scale (CPGS), Short Form-36 Bodily Pain Scale (SF-36 BPS), and Measure of Intermittent and Constant Osteoarthritis Pain (ICOAP). Arthritis Care Res 63(Suppl 11):S240-S252, 2011

40. Hazard RG, Haugh LD, Green PA, Jones PL: Chronic low back pain: The relationship between patient satisfaction and pain, impairment, and disability outcomes. Spine 19:881-887, 1994

41. Henry SG, Bell RA, Fenton JJ, Kravitz RL: Goals of chronic pain management: Do patients and primary care physicians agree and does it matter? Clin J Pain 33:955-961, 2017

42. Hjermstad MJ, Fayers PM, Haugen DF, Caraceni A, Hanks GW, Loge JH, Fainsinger R, Aass N, Kaasa S: Studies comparing numerical rating scales, verbal rating scales, and visual analogue scales for assessment of pain intensity in adults: A systematic literature review. J Pain Symptom Manage 41:1073-1093, 2011

43. Hush JM, Refshauge K, Sullivan G, De Souza L, Maher CG, McAuley JH: Recovery: What does this mean to patients with low back pain? Arthritis Rheum 61:124-131, 2009

44. Hush JM, Refshauge KM, Sullivan G, De Souza L, McAuley JH: Do numerical rating scales and the Roland-Morris Disability Questionnaire capture changes that are meaningful to patients with persistent back pain? Clin Rehabil 24:648-657, 2010

45. Huskisson E: Measurement of pain. Lancet 304:1127-1131, 1974

46. Jamison RN, Raymond SA, Slawsby EA, McHugo GJ, Baird JC: Pain assessment in patients with low back pain: Comparison of weekly recall and momentary electronic data. J Pain 7:192-199, 2006

47. Jensen MP: The validity and reliability of pain measures for use in clinical trials in adults: Review paper written for the Initiative on Methods, Measurements, and Pain Assessment in Clinical Trials (IMMPACT) meeting. April 12−13 (2003), IMMPACT-II, http://www.immpact.org/static/meetings/Immpact2/background/Jensen_review.pdf, 2003. Accessed November 3, 2017

48. Jensen MP, Hu X, Potts SL, Gould EM: Single vs composite measures of pain intensity: Relative sensitivity for detecting treatment effects. Pain 154:534-538, 2013

49. Jensen MP, Schnitzer TJ, Wang H, Smugar SS, Peloso PM, Gammaitoni A: Sensitivity of single-domain versus multiple-domain outcome measures to identify responders in chronic low-back pain: Pooled analysis of 2 placebo-controlled trials of etoricoxib. Clin J Pain 28:1-7, 2012

50. Jensen MP, Tome-Pires C, Sole E, Racine M, Castarlenas E, de la Vega R, Miro J: Assessment of pain intensity in clinical trials: Individual ratings vs composite scores. Pain Med 16:141-148, 2015

51. Jensen MP, Turner JA, Romano JM, Fisher LD: Comparative reliability and validity of chronic pain intensity measures. Pain 83:157-162, 1999

52. Kahl C, Cleland JA: Visual analogue scale, numeric pain rating scale and the McGill Pain Questionnaire: An overview of psychometric properties. Phys Ther Rev 10:123-128, 2005

53. Kaiser U, Kopkow C, Deckert S, Neustadt K, Jacobi L, Cameron P, De VA, Apfelbacher C, Arnold B, Birch J: Developing a core outcome-domain set to assessing effectiveness of interdisciplinary multimodal pain therapy: The VAPAIN consensus statement on core outcome-domains. Pain 159:673-683, 2018

54. Kamper SJ, Apeldoorn AT, Chiarotto A, Smeets RJ, Ostelo RW, Guzman J, van Tulder MW: Multidisciplinary biopsychosocial rehabilitation for chronic low back pain. Cochrane Database Syst Rev , 2014:Cd000963

55. Keller S, Bann CM, Dodd SL, Schein J, Mendoza TR, Cleeland CS: Validity of the brief pain inventory for use in documenting the outcomes of patients with noncancer pain. Clin J Pain 20:309-318, 2004

56. Kovacs FM, Abraira V, Royuela A, Corcoll J, Alegre L, Cano A, Muriel A, Zamora J, del Real MT, Gestoso M, Mufraggi N: Minimal clinically important change for pain intensity and disability in patients with nonspecific low back pain. Spine 32:2915-2920, 2007

57. Krebs EE, Bair MJ, Damush TM, Tu W, Wu J, Kroenke K: Comparative responsiveness of pain outcome measures among primary care patients with musculoskeletal pain. Med Care 48:1007-1014, 2010

58. Lapane KL, Quilliam BJ, Benson C, Chow W, Kim M: One, two, or three? Constructs of the brief pain inventory among patients with non-cancer pain in the outpatient setting. J Pain Symtpom Manage 47:325-333, 2014

59. Lauridsen HH, Hartvigsen J, Manniche C, Korsholm L, Grunnet-Nilsson N: Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients. BMC Musculoskelet Disord 7:82, 2006

60. Lauridsen HH, Manniche C, Korsholm L, Grunnet-Nilsson N, Hartvigsen J: What is an acceptable outcome of treatment before it begins? Methodological considerations and implications for patients with chronic low back pain. Eur Spine J 18:1858-1866, 2009

61. Law M, McIntosh J, Morrison L, Baptiste S: A comparison of two pain measurement scales: Their clinical value. Can J Rehabil 1:55-58, 1987

62. Lee H, Hübscher M, Moseley GL, Kamper SJ, Traeger AC, Mansell G, McAuley JH: How does pain lead to disability? A systematic review and meta-analysis of mediation studies in people with back and neck pain. Pain 156:988-997, 2015

63. Lee TH: Zero pain is not the goal. JAMA 315:1575-1577, 2016

64. Love A, Leboeuf C, Crisp TC: Chiropractic chronic low back pain sufferers and self-report assessment methods. Part I. A reliability study of the visual analogue scale, the pain drawing and the McGill Pain Questionnaire. J Manipulative Physiol Ther 12:21-25, 1989

65. Machado GC, Maher CG, Ferreira PH, Day RO, Pinheiro MB, Ferreira ML: Non-steroidal anti-inflammatory drugs for spinal pain: A systematic review and meta-analysis. Ann Rheum Dis 76:1269-1278, 2017

66. Machin D, Lewith GT, Wylson S: Pain measurement in randomized clinical trials: A comparison of two pain scales. Clin J Pain 4:161-168, 1988

67. Maher C, Underwood M, Buchbinder R: Non-specific low back pain. Lancet 389:736-747, 2017

68. Marin TJ, Van Eerd D, Irvin E, Couban R, Koes BW, Malmivaara A, van Tulder MW, Kamper SJ: Multidisciplinary biopsychosocial rehabilitation for subacute low back pain. Cochrane Database Syst Rev 6, 2017:CD002193

69. Maughan EF, Lewis JS: Outcome measures in chronic low back pain. Eur Spine J 19:1484-1494, 2010

70. McRae M, Hancock MJ: Adults attending private physiotherapy practices seek diagnosis, pain relief, improved function, education and prevention: A survey. J Physiother 63:250-256, 2017

71. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P: Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Med 6, 2009:e1000097

72. Mokkink LB, De Vet HC, Prinsen CA, Patrick DL, Alonso J, Bouter LM, Terwee CB: COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. Qual Life Res 27:1171-1179, 2018

73. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC: The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 63:737-745, 2010

74. Olaogun MO, Adedoyin RA, Ikem IC, Anifaloba OR: Reliability of rating low back pain with a visual analogue scale and a semantic differential scale. Physiother Theory Pract 20:135-142, 2004

75. Ostelo RW, Deyo RA, Stratford P, Waddell G, Croft P, Von Korff M, Bouter LM, de Vet HC: Interpreting change scores for pain and functional status in low back pain: Towards international consensus regarding minimal important change. Spine 33:90-94, 2008

76. Ostelo RW, Swinkels-Meewisse IJ, Knol DL, Vlaeyen JW, de Vet HC: Assessing pain and pain-related fear in acute low back pain: What is the smallest detectable change? Int J Behav Med 14:242-248, 2007

77. Page MJ, Huang H, Verhagen AP, Buchbinder R, Gagnier JJ: Identifying a core set of outcome domains to measure in clinical trials for shoulder disorders: A modified Delphi study. RMD Open 2, 2016:e000380

78. Park KB, Shin JS, Lee J, Lee YJ, Kim MR, Lee JH, Shin KM, Shin BC, Cho JH, Ha IH: Minimum clinically important difference and substantial clinical benefit in pain, functional, and quality of life scales in failed back surgery syndrome patients. Spine 42:E474-e481, 2017

79. Parreira P, Heymans MW, van Tulder MW, Esmail R, Koes BW, Poquet N, Lin CWC, Maher CG: Back schools for chronic non-specific low back pain. Cochrane Database Syst Rev 8, 2017:CD011674

80. Paungmali A, Sitilertpisan P, Taneyhill K, Pirunsan U, Uthaikhup S: Intrarater reliability of pain intensity, tissue blood flow, thermal pain threshold, pressure pain threshold and lumbo-pelvic stability tests in subjects with low back pain. Asian J Sports Med 3:8-14, 2012

81. Pengel LH, Refshauge KM, Maher CG: Responsiveness of pain, disability, and physical impairment outcomes in patients with low back pain. Spine 29:879-883, 2004

82. Price DD, Harkins SW: Combined use of experimental pain and visual analogue scales in providing standardized measurement of clinical pain. Clin J Pain 3:1-8, 1987

83. Price DD, McGrath PA, Rafii A, Buckingham B: The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. Pain 17:45-56, 1983

84. Prinsen CA, Mokkink LB, Bouter LM, Alonso J, Patrick DL, De Vet HC, Terwee CB: COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res 27:1147-1157, 2018

85. Prinsen CA, Vohra S, Rose MR, Boers M, Tugwell P, Clarke M, Williamson PR, Terwee CB: How to select outcome measurement instruments for outcomes included in a “Core Outcome Set”: A practical guideline. Trials 17:449, 2016

86. Ramasamy A, Martin ML, Blum SI, Liedgens H, Argoff C, Freynhagen R, Wallace M, McCarrier KP, Bushnell DM, Hatley NV, Patrick DL: Assessment of patient-reported outcome instruments to assess chronic low back pain. Pain Med 18:1098-1110, 2017

87. Roach KE, Brown MD, Dunigan KM, Kusek CL, Walas M: Test-retest reliability of patient reports of low back pain. J Orthop Sports Phys Ther 26:253-259, 1997

88. Robinson-Papp J, George MC, Dorfman D, Simpson DM: Barriers to chronic pain measurement: A qualitative study of patient perspectives. Pain Med 16:1256-1264, 2015

89. Scrimshaw SV, Maher C: Responsiveness of visual analogue and McGill pain scale measures. J Manipulative Physiol Ther 24:501-504, 2001

90. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, Porter AC, Tugwell P, Moher D, Bouter LM: Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 7:10, 2007

91. Sheldon EA, Bird SR, Smugar SS, Tershakovec AM: Correlation of measures of pain, function, and overall response: Results pooled from two identical studies of etoricoxib in chronic low back pain. Spine 33:533-538, 2008

92. Solomon D, Roopchand-Martin S, Swaminathan N, Heymans MW: How well do pain scales correlate with each other and with the Oswestry Disability Questionnaire? Int J Ther Rehabil 18, 2011

93. Song CY, Lin SF, Huang CY, Wu HC, Chen CH, Hsieh CL: Validation of the Brief Pain Inventory in patients with low back pain. Spine 41:E937-E942, 2016

94. Stinson JN, Kavanagh T, Yamada J, Gill N, Stevens B: Systematic review of the psychometric properties, interpretability and feasibility of self-report pain intensity measures for use in clinical trials in children and adolescents. Pain 125:143-157, 2006

95. Strong J, Ashton R, Chant D: Pain intensity measurement in chronic low back pain. Clin J Pain 7:209-218, 1991

96. Sullivan MD, Ballantyne JC: Must we reduce pain intensity to treat chronic pain? Pain 157:65-69, 2016

97. Tan G, Jensen MP, Thornby JI, Shanti BF: Validation of the Brief Pain Inventory for chronic nonmalignant pain. J Pain 5:133-137, 2004

98. Terwee CB, Jansma EP, Riphagen II, de Vet HC: Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res 18:1115-1123, 2009

99. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC: 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, 2012

100. Terwee CB, Prinsen CA, Chiarotto A, De Vet HC, Westerman MJ, Patrick DL, Alonso J, Bouter LM, Mokkink LB: COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. Qual Life Res 27:1159-1170, 2018

101. Triano JJ, McGregor M, Cramer GD, Emde DL: A comparison of outcome measures for use with back pain patients: Results of a feasibility study. J Manipulative Physiol Ther 16:67-73, 1993

102. Turk DC, Dworkin RH, Allen RR, Bellamy N, Brandenburg N, Carr DB, Cleeland C, Dionne R, Farrar JT, Galer BS, Hewitt DJ, Jadad AR, Katz NP, Kramer LD, Manning DC, McCormick CG, McDermott MP, McGrath P, Quessy S, Rappaport BA, Robinson JP, Royal MA, Simon L, Stauffer JW, Stein W, Tollett J, Witter J: Core outcome domains for chronic pain clinical trials: IMMPACT recommendations. Pain 106:337-345, 2003

103. Turk DC, Dworkin RH, Revicki D, Harding G, Burke LB, Cella D, Cleeland CS, Cowan P, Farrar JT, Hertz S, Max MB, Rappaport BA: Identifying important outcome domains for chronic pain clinical trials: An IMMPACT survey of people with pain. Pain 137:276-285, 2008

104. van der Roer N, Ostelo RW, Bekkering GE, van Tulder MW, de Vet HC: Minimal clinically important change for pain intensity, functional status, and general health status in patients with nonspecific low back pain. Spine 31:578-582, 2006

105. Whiting P, Savovic J, Higgins JP, Caldwell DM, Reeves BC, Shea B, Davies P, Kleijnen J, Churchill R: ROBIS: A new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol 69:225-234, 2016

106. Whynes DK, McCahon RA, Ravenscroft A, Hodgkinson V, Evley R, Hardman JG: Responsiveness of the EQ-5D health-related quality-of-life instrument in assessing low back pain. Value Health 16:124-132, 2013

107. Williams ACdC, Davies HTO, Chadury Y: Simple pain rating scales hide complex idiosyncratic meanings. Pain 85:457-463, 2000

108. Williamson A, Hoggart B: Pain: A review of three commonly used pain rating scales. J Clin Nurs 14:798-804, 2005

109. Wright AA, Cook CE: Criterion validation of the rate of recovery, single alphanumeric measure, in patients with low back pain. Physiother Res Int 18:124-129, 2013