Measurement properties of translated versions of the Scoliosis Research Society-22 Patient Questionnaire, SRS-22: a systematic review

Marco Monticone · Claudia Nava · Vittorio Leggero · Barbara Rocca · Stefano Salvaderi · Simona Ferrante · Emilia Ambrosini

Accepted: 2 February 2015

M. Monticone (✉) · C. Nava · V. Leggero · B. Rocca · S. Salvaderi · E. Ambrosini
Physical Medicine and Rehabilitation Unit, Scientific Institute of Lissone, Institute of Care and Research, Salvatore Maugeri Foundation, IRCCS, Via Monsignor Bernasconi 16, 20035 Lissone, Monza Brianza, MI, Italy
e-mail: marco.monticone@fsm.it

S. Ferrante · E. Ambrosini
Neuroengineering and Medical Robotics Laboratory, Department of Electronics, Informations and Bioengineering, Politecnico di Milano, Milan, Italy

Abbreviations

AIS	Adolescent idiopathic scoliosis
AUC	Area under the curve
CHQ-CF87	Child Health Questionnaire-Child Form 87
COSMIN	COnsensus-based Standards for the selection of health status Measurement INstruments
DIF	Differential item functioning
HRQoL	Health-Related Quality of Life
ICC	Intraclass correlation coefficient

IRT	Item response theory
LOA	Limits of agreement
MIC	Minimal important change
NCSS	Non-clinically significant scoliosis
RMDQ	Roland and Morris Disability Questionnaire
SDC	Smallest detectable change
SF-12	Short-Form Health Survey-12 items
SF-36	Short-Form Health Survey-36 items
SR	Systematic review
SRS-22	Scoliosis Research Society-22 Patient Questionnaire
VAS	Visual analogue scale

Introduction

The region-specific Scoliosis Research Society-22 Patient Questionnaire (SRS-22) was developed in English and recommended by Asher in 2003 as a means of assessing Health-Related Quality of Life (HRQoL) in subjects with scoliosis [1, 2]. The SRS-22, reported in Appendix 1, is multidimensional and covers five domains named Function (evaluating current level of activity and motor performances of the spine during usual activities at home, at school, at work, etc), Pain (investigating painful sensations felt during the past months, drugs eventually used, and sick days due to back pain), Mental Health (evaluating mental sensations such as anxiety, depression, peace of mind, sadness, and happiness), Self-image (investigating esthetic aspect and self-appeal), and Management Satisfaction/Dissatisfaction (assessing patients' satisfaction with treatment results). Each domain has five questions except the last, which has two. The total score for each item ranges from 1 to 5, with 5 being the best. Each domain has a total sum score ranging from 5 to 25, except for the satisfaction domain, which ranges from 2 to 10. The sum of the first four domains gives a maximum subtotal of 100; with addition of the satisfaction domain, the maximum total is 110. Results are usually expressed as the mean (total sum of the domain divided by the number of items answered) for each domain. The SRS-22 is expected to allow a comprehensive evaluation of the disease and subjects' perceptions of the consequences of clinician choices and the effectiveness of treatments [3–6].

Many efforts were made in order to adapt and investigate its properties also in non-English countries, with the purpose of allowing clinicians and researchers to share validated outcomes and start high-quality methodological trials in the field of scoliosis. Over the last decade, the amount of studies investigating the measurement properties of translations of the SRS-22 increased considerably, but, to the best of our knowledge, no systematic review (SR) has ever been made of their psychometric properties.

The aim of this review was to evaluate the psychometric properties and to provide the current level of evidence of all the available translations of the SRS-22 using the "COnsensus-based Standards for the selection of health status Measurement INstruments" (COSMIN) [7]. Specific targets were as follows:

- to examine the methodological quality of the studies examining psychometric properties;
- to rate the quality of the SRS-22 translations in terms of psychometric properties;
- to provide an overall level of evidence of the measurement properties per language.

Methods

The study was approved by Institutional Review Board of the Salvatore Maugeri Foundation's Scientific Institute in Lissone.

Search strategy and selection criteria

We searched the PubMed, Medline, EMbase, and CINAHL databases for articles published up to January 2014. The combination of the following terms and their derivates were searched in each database: "Scoliosis Research Society-22," "validation," "transcultural adaptation," "psychometric properties," "quality of life," "outcome measure" (see Appendix 2 for the full search strategy). All of the possible keywords used in each database were included in the search. The reference lists of the selected articles were also screened to identify additional studies.

The search considered original, full-text articles published in English describing translations of the SRS-22 or evaluating their measurement properties. Scoliosis had to be the main problem of investigated subjects.

Two reviewers (CN and VL) independently assessed the titles, abstracts, and reference lists of the studies retrieved during the literature search. In the case of disagreement, a discussion was held in an attempt to reach a consensus, and, if necessary, a third reviewer (MM) made the final decision.

Assessment of the methodological quality

The studies were assessed using the COSMIN checklist, which is a standardized and validated scoring tool developed in an international and multidisciplinary Delphi study [8–11]. It focuses on standards for design requirements and preferred statistical methods of studies, which investigate the measurement properties of health measurement instruments. It consists of 114 items grouped in twelve boxes:

nine contain standards for measurement properties, one standards for interpretability, and two the requirements for item response theory (IRT) and the generalizability of the results. Each item was scored on a 4-point rating scale ("poor," "fair," "good," or "excellent") [10], and the lowest rating of the items in a box was used to define the overall methodological quality of each study.

The data were extracted independently by two reviewers (CN and VL), who also assessed their methodological quality. In the case of disagreement, a discussion was held in an attempt to reach a consensus, and, if necessary, a third reviewer (MM) made the final decision. In order to decrease the differences between reviewers, a scoring system was agreed in advance. The data extraction form was designed and tested before being used for the purposes of the study.

Rating of the psychometric properties

In order to rate the psychometric properties of each translation, we applied a validated quality assessment criteria proposed by Terwee et al. [12]. The properties were rated "positive," "negative," or "indeterminate" according to the examination of the results of the measurement properties. Concerning properties which required an independent evaluation for each subscales (i.e., Cronbach's α, ICC, MIC, Pearson's coefficient), we considered the rating as "positive," "negative," or "indeterminate" if the criterion was satisfied in at least 4 subscales out of 5. Appendix 3 lists the quality criteria for each of the properties considered.

Best evidence synthesis

An overall level of evidence of the measurement properties per language was determined combining scores of the methodological quality assessment (COSMIN checklist) and the psychometric results rated with Terwee's classification [12]. This method of evaluation was developed by the Cochrane Back Review Group [13, 14] and consists of five possible levels of evidence: "strong," "moderate," "limited," "conflicting," or "unknown" (Table 1).

Table 1 Levels of evidence for the overall quality of the measurement properties (based on the Cochrane Back Review Group) [13, 14]

Level	Rating[a]	Criteria
Strong evidence	+++ or −−−	Consistent findings in multiple studies of good methodological quality or in one study of excellent methodological quality
Moderate evidence	++ or −−	Consistent findings in multiple studies of fair methodological quality or in one study of good methodological quality
Limited evidence	+ or −	One study of fair methodological quality
Conflicting evidence	±	Conflicting findings
Unknown evidence	?	Only studies of poor methodological quality

[a] +, means positive rating; −, means negative rating; ±, means conflicting rating; ?, means indeterminate rating

Measurement properties, interpretability, and generalizability

Measurement properties were divided into three domains [11]: The first included internal consistency (i.e., the interrelatedness of the items), measurement error (i.e., systematic and random errors in patient scores that are not attributable to true changes in the construct being measured), and reliability (i.e., the proportion of total variance due to "true" differences between patients); the second included content validity (i.e., the extent to which the content adequately reflects the construct being measured), criterion validity (i.e., the extent to which the scores adequately reflect a gold standard), construct validity divided into cross-cultural validity (i.e., the extent to which the performance of an item in the adapted instrument adequately reflects its performance in the original version) and structural validity (i.e., the extent to which the scores adequately reflect the dimensional nature of the construct being measured), and hypotheses testing (i.e., the extent to which the measure being evaluated relates to other measures in the expected manner); and the third covered responsiveness (i.e., the instrument's ability to detect changes over time in the construct being measured), which is considered an aspect of longitudinal validity and assessed using the same standards as those used for the other aspects of validity [9].

Interpretability (i.e., the extent to qualitative meaning can be attributed to the instrument's quantitative scores or changes in scores) was assessed by considering the clinically relevant differences in scores between subgroups, floor and ceiling effects, and minimal important changes [9]. Ceiling and floor effects were detected when almost 15 % of respondents achieved the highest/lowest possible score [12]. The generalizability of each property (i.e., the extent to which the results can be generalized) was assessed by gathering information concerning age, gender, disease characteristics, setting, and subject selection methods.

Results

The search strategy identified 106 references in PubMed database, 54 references in Medline, 41 references in CINAHL, and 47 references in EMbase. After screening for duplicates, a total of 121 articles were found, 98 of which

were selected on the basis of their titles and abstracts. The subsequent full-text assessment excluded 74 articles with the following reasons: The SRS-22 was not the main purpose of the study ($n = 39$), the evaluation of the measurement properties of SRS-22 was not the main aim of the study ($n = 20$), the article was not written in English ($n = 2$), and the study referred to other versions of the SRS questionnaire ($n = 13$). Therefore, the review was performed on 24 articles concerning the psychometric properties of the SRS-22 into 17 different languages (Fig. 1).

Reviewers independently assessed and extracted data about the methodological quality of each study by using the COSMIN checklist. A degree of consensus of about 90 % was reached between reviewers. Disagreement was solved by discussion between reviewers in 6 % of cases, while the third reviewer was involved in order to make the final decision in 4 % of cases of disagreements.

Table 2 shows the general characteristics of the included studies.

Table 3 reports the methodological quality of each translation and property. No studies performed multigroup factor analysis or differential item functioning; hence, we were only able to rate the methodological quality of the translation process. Internal consistency, reliability, and hypotheses testing were the properties mainly investigated, and in the majority of cases, the methodological quality was fair.

Table 4 summarizes the outcomes of the included studies and the corresponding ratings for each investigated measurement property based on the quality criteria proposed by Terwee et al. [12]. The overall assessment of the investigated measurements properties was positive (74 %), negative (25 %), and indeterminate (3 %). Table 4 reports also information about interpretability related to

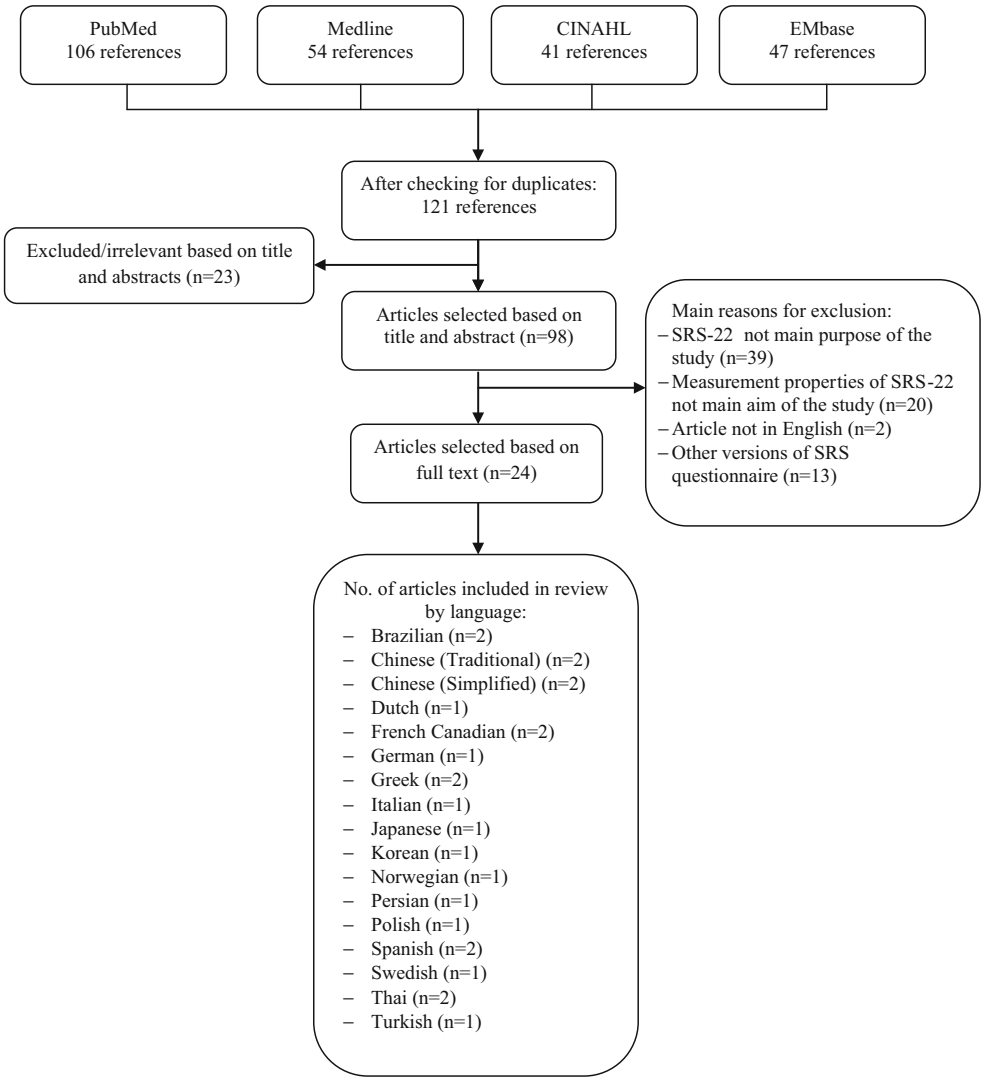

Fig. 1 Flowchart search and selection

Table 2 General characteristics of the considered studies

Study	Language	Country	Population	Treatment	Setting
Rosanova et al. [15]	Brazilian	Brazil	$n = 49$ AIS; $M = 7$, $F = 47$; 19.9 ± 7.7 years	Post-surgery and conservative treatment	Orthopedic
Camarini et al. [16]	Brazilian	Brazil	$n = 44$ AIS; $M = 4$, $F = 40$; 18.9 (12–36) years	Post-surgery and conservative treatment	Orthopedic
Zhao et al. [17]	Chinese (traditional)	China	$n = 86$ AIS; $M = 11$, $F = 75$; 13.9 (10–18) years	Conservative treatment	Orthopedic
Cheung et al. [18]	Chinese (traditional)	China	$n = 36$ AIS; $M = 4$, $F = 32$; 16.5 (8–28) years; $n = 50$ AIS; $M = 4$, $F = 46$; 21 (12–51) years	Not reported	Outpatient
Li et al. [19]	Chinese (simplified)	China	$n = 63$ AIS; $M = 6$, $F = 57$; 17.7 (14.3–23.8) years	Post-surgery	Home-based
Qiu et al. [20]	Chinese (simplified)	China	$n = 333$ AIS; $M = 76$, $F = 257$; 16.2 (10–32) years	Pre-/post-surgery; Before/after brace	Orthopedic
Schlösser et al. [21]	Dutch	The Netherlands	$n = 92$ AIS; $M = 12$, $F = 80$; 15.1 ± 2.0 years	Post-surgery, conservative treatment, observation	Outpatient
Beauséjour et al. [22]	French Canadian	Canada	$n = 145$ AIS; $M = 22$, $F = 123$; 15 (9.8–21.2) years; $n = 44$ NCSS; $M = 13$, $F = 31$; 13.5 (9.8–21.6) years; $n = 64$ healthy subjects; $M = 23$, $F = 41$; 14.1 (10.3–17.8) years	Not reported	Outpatient
Lonjon et al. [23]	French Canadian	France	$n = 175$ AIS; $M = 32$, $F = 143$; 14.4 (10.1– 18.9) years; $n = 25$ NCSS; $M = 5$, $F = 20$; 13.8 (12–18.6) years; $n = 60$ healthy subjects; $M = 13$, $F = 47$; 14.7 (12.7–18.3) years	Conservative treatment or surgery scheduled (AIS group); Observation (NCSS group)	Outpatient orthopedic
Niemeyer et al. [24]	German	Germany	$n = 78$ AIS; $M = 18$, $F = 60$; 19 (14–59) years	Post-surgery and conservative treatment	Home-based
Antonarakos et al. [25]	Greek	Greece	$n = 51$ AIS; $M =$ NA, $F =$ NA; 21.2 (16–27) years	Post-surgery	Home-based

Table 2 continued

Study	Language	Country	Population	Treatment	Setting
Potoupnis et al. [26]	Greek	Greece	$n = 87$ AIS; $M = 7$, $F = 80$; 14.78 (12–18) years	Conservative treatment	Home-based
Monticone et al. [27]	Italian	Italy	$n = 223$ AIS; $M = 103$, $F = 120$; 14.14 (10–18) years	Conservative treatment	Rehabilitation
Hashimoto et al. [28]	Japanese	Japan	$n = 114$ AIS; $M = 9$, $F = 105$; 15 (11–17.9) years	Post-surgery, conservative treatment and observation	Orthopedic
Lee et al. [29]	Korean	Republic of Korea	$n = 82$ AIS; $M = 12$, $F = 70$; 18.3 (14.1–24.2) years	Post-surgery	Home-based
Adobor et al. [30]	Norwegian	Norway	$n = 57$; $M = 9$, $F = 48$; 21 (12–45) years	Pre-/post-surgery; Brace	Home-based
Mousavi et al. [31]	Persian	Iran	$n = 84$ AIS; $M = 25$, $F = 59$; 15.32 (12–18) years	Post-surgery, conservative treatment and observation	Orthopedic
Glowacki et al. [32]	Polish	Poland	$n = 60$ AIS; $M = 0$, $F = 60$; 15.6 (12–28) years	Post-surgery	Orthopedic
Bago et al. [33]	Spanish	Spain	$n = 175$ AIS; $M = 23$, $F = 152$; 18.9 (8–48) years	Post-surgery, conservative treatment and observation	Orthopedic
Climent et al. [34]	Spanish	Spain	$n = 175$ AIS; $M = 23$, $F = 152$; 18.9 (8–48) years	Post-surgery, conservative treatment, observation	Orthopedic
Danielsson et al. [35]	Swedish	Sweden	$n = 193$ AIS; $M = 13$, $F = 180$; 23.3 (11.9–56.9) years	Post-surgery, conservative treatment, observation	Orthopedic
Leelapattana et al. [36]	Thai	Thailand	$n = 30$ AIS; $M = 2$, $F = 28$; 17.2 (13–30) years	Post-surgery, conservative treatment	Orthopedic
Sathira-Angkura et al. [37]	Thai	Thailand	$n = 58$ AIS; $M = 52$, $F = 6$; 18.7 (13–31) years	Post-surgery	General hospital
Alanay et al. [38]	Turkish	Turkey	$n = 47$ AIS; $M = 12$, $F = 35$; 19.8 (14–31) years	Post-surgery	Home-based

AIS adolescent idiopathic scoliosis, *M* male, *F* female, *NCSS* non-clinically significant scoliosis, *NA* not available

floor/ceiling effects: Function had a ceiling effect in 10 studies, Pain in 18, Self-image in 2, Mental Health in 3, and Management Satisfaction/Dissatisfaction in 12. The studies that considered floor effects ($n = 20$) did not observe any. None of the studies used IRT methods.

The overall level of evidence of each measurement property per language is reported in Table 5. Internal consistency was mainly judged as positive with a strong/moderate to limited level of evidence. Measurement error was assessed only in one translation and achieved a limited negative evidence. Reliability reported limited positive evidence or unknown evidence due to poor methodological quality. Structural validity was evaluated in only six languages, achieving half positive

Table 3 Methodological quality of each study by measurement property

Language	Cross-cultural validity	Internal consistency	Measurement error	Reliability	Structural validity	Hypothesis testing
Brazilian						
Rosanova et al. [15]	NA	NA	NA	NA	NA	Fair
Camarini et al. [16]	Poor	Fair	NA	Fair	NA	NA
Chinese (traditional)						
Zhao et al. [17]	Poor	Fair	NA	Poor	Fair	NA
Cheung et al. [18]	Poor	Fair	NA	Fair	NA	Fair
Chinese (simplified)						
Li et al. [19]	Poor	Fair	NA	Poor	NA	Fair
Qiu et al. [20]	Poor	Fair	NA	Poor	NA	Fair
Dutch						
Schlösser et al. [21]	Excellent	Fair	NA	Fair	NA	Fair
French Canadian						
Beauséjour et al. [22]	Fair	Fair	NA	NA	Fair	Fair
Lonjon et al. [23]	NA	Fair	Fair	Fair	Fair	Fair
German						
Niemeyer et al. [24]	Fair	Fair	NA	Poor	NA	Fair
Greek						
Antonarakos et al. [25]	Fair	Fair	NA	Poor	NA	Fair
Potoupnis et al. [26]	Poor	Fair	NA	Fair	Fair	Fair
Italian						
Monticone et al. [27]	Excellent	Excellent	NA	Fair	Excellent	Fair
Japanese						
Hashimoto et al. [28]	Poor	Fair	NA	NA	Fair	Fair
Korean						
Lee et al. [29]	Excellent	Fair	NA	Poor	NA	Fair
Norwegian						
Adobor et al. [30]	Poor	Fair	NA	Fair	NA	Fair
Persian						
Mousavi et al. [31]	Fair	Fair	NA	Poor	NA	Fair
Polish						
Glowackiet al. [32]	Poor	Fair	NA	NA	NA	NA
Spanish						
Bago et al. [33]	Poor	Fair	NA	Fair	NA	NA
Climent et al. [34]	NA	NA	NA	NA	Good	Fair
Swedish						
Danielsson et al. [35]	Excellent	Fair	NA	Poor	NA	Fair
Thai						
Leelapattana et al. [36]	Poor	Fair	NA	Poor	NA	Fair
Sathira-Angkura et al. [37]	Fair	Fair	NA	Fair	NA	Fair
Turkish						
Alanay et al. [38]	Poor	Fair	NA	Poor	NA	Fair

Content validity, criterion validity, and responsiveness are not reported in the table, since no studies included in the review evaluated these measurement properties

NA (not available) means that the psychometric property was not investigated

and half negative or unknown results. Hypotheses testing had a limited positive evidence in eight languages, conflicting evidence in three, and limited or moderate negative evidence in the remaining five.

Content validity, criterion validity, and responsiveness are not reported in Tables 3, 4 and 5, since no studies included in the review evaluated these measurement properties.

Results per language are detailed below.

Table 4 Outcomes and corresponding ratings for each measurement properties

Language		Internal consistency (Cronbach's α)	Measurement error (SEM)	Reliability (test/retest time intervals and intraclass coefficient correlations by subscale)	Structural Validity (% of explained variance)	Hypotheses Testing[a] (measure comparisons and Pearson/Spearman correlation coefficients)	Floor effects (%)	Ceiling effects (%)
Brazilian								
Rosanova et al. [15]	Study results	NA	NA	NA	NA	SRS-22 and SF-36; $F = 0.49–0.83$ $P = 0.69–0.86$ SI = 0.34–0.40 MH = 0.39–0.57 $S = 0.07–0.28$	NA	NA
	Score[b]	NA	NA	NA	NA	−		
Camarini et al. [16]	Study results	$F = 0.77$ $P = 0.80$ SI = 0.82 MH = 0.85 $S = 0.70$	NA	7 days; $F = 0.94$ $P = 0.93$ SI = 0.92 MH = 0.92 $S = 0.96$	NA	NA	$F = 0$ $P = 0$ SI = 0 MH = 0 $S = 0$	$F = 15.9$ $P = 25$ SI = 4.54 MH = 2.27 $S = 36.36$
	Score[b]	+	NA	+	NA	NA		
Chinese (traditional)								
Zhao et al. [17]	Study results	$F = 0.70$ $P = 0.80$ SI = 0.80 MH = 0.88 $S = 0.81$	NA	3–4 weeks; $F = 0.85$ $P = 0.96$ SI = 0.96 ME = 0.95 $S = 0.91$	67.66	NA	$F = 2.3$ $P = 7$ SI = 2.3 MH = 4.7 $S = 2.3$	$F = 2.3$ $P = 15.1$ SI = 4.7 MH = 5.8 $S = 2.3$
	Score[b]	+	NA	+	+	NA		
Cheung et al. [18]	Study results	$F = 0.86$ $P = 0.87$ SI = 0.78 MH = 0.87 $S = 0.53$	NA	7 days; $F = 0.83$ $P = 0.76$ SI = 0.79 MH = 0.84 $S = 0.82$	NA	SRS-22 and SF-36; $F = 0.59–0.77$ $P = 0.54–0.72$ SI = 0.50–0.62 MH = 0.57–0.67 $S = 0.18–0.49$	$F = 2$ $P = 2$ SI = 4 MH = 4 $S = 2$	$F = 44$ $P = 30$ SI = 2 MH = 18 $S = 10$
	Score[b]	+	NA	+	NA	+		
Chinese (simplified)								
Li et al. [19]	Study results	$F = 0.81$ $P = 0.88$ SI = 0.76 MH = 0.79 $S = 0.65$	NA	13, 4 days; $F = 0.74$ $P = 0.78$ SI = 0.86 MH = 0.81 $S = 0.84$	NA	SRS-22 and SF-36; $F = 0.50–0.76$ $P = 0.46–0.81$ SI = 0.37–0.62 MH = 0.60–0.85 $S = 0.23–0.44$	$F = 1.6$ $P = 1.6$ SI = 1.6 MH = 1.6 $S = 3.2$	$F = 4.8$ $P = 22.2$ SI = 1.6 MH = 9.5 $S = 14.3$
	Score[b]	+	NA	+	NA	+		
Qiu et al. [20]	Study results	$F = 0.57$ $P = 0.73$ SI = 0.71 MH = 0.79 $S = 0.50$	NA	12 days; $F = 0.78$ $P = 0.70$ SI = 0.85 MH = 0.82 $S = 0.75$	NA	SRS-22 and SF-36; $F = 0.49–0.70$ $P = 0.33–0.52$ SI = 0.24–0.63 MH = 0.42–0.62 $S = 0.21–0.34$	$F = 5.1$ $P = 1.2$ SI = 5.9 MH = 1.2 $S = 2.2$	$F = 36.4$ $P = 67.5$ SI = 22.6 MH = 26.6 $S = 14.0$
	Score[b]	−	NA	+	NA	−		

Table 4 continued

Language		Internal consistency (Cronbach's α)	Measurement error (SEM)	Reliability (test/retest time intervals and intraclass coefficient correlations by subscale)	Structural Validity (% of explained variance)	Hypotheses Testing[a] (measure comparisons and Pearson/Spearman correlation coefficients)	Floor effects (%)	Ceiling effects (%)
Dutch								
Schlösser et al. [21]	Study results	$F = 0.74$ $P = 0.85$ SI = 0.71 MH = 0.77 $S = 0.71$	NA	19 days; $F = 0.86$ $P = 0.92$ SI = 0.87 MH = 0.85 $S = 0.79$	NA	SRS-22 and SF-36; $F = 0.42$–0.80 $P = 0.69$–0.85 SI = 0.39–0.49 MH = 0.45–0.59	$F = 0$ $P = 1$ SI = 0 MH = 0 $S = 0$	$F = 33$ $P = 20$ SI = 4 MH = 8 $S = 22$
	Score[b]	+	NA	+	NA	+		
French Canadian								
Beauséjour et al. [22]	Study results	$F = 0.68$ $P = 0.79$ SI = 0.67 MH = 0.79 $S = 0.69$ (AIS patients group)	NA	NA	47.4	SRS-22 and SF-12; $F = 0.46$–0.64 $P = 0.62$–0.75 SI = 0.23–0.50 MH = 0.47–0.49 $S = 0.18$–0.26 (AIS patients group)	$F = 0$ $P = 0$ SI = 0 MH = 0 $S = 0.7$ (AIS patients group)	$F = 3.4$ $P = 22.1$ SI = 6.9 MH = 0.7 $S = 22.1$ (AIS patients group)
	Score[b]	−	NA	NA	−	−		
Lonjon et al. [23]	Study results	Only AIS patients group $F = 0.60$ $P = 0.71$ SI = 0.61 MH = 0.73 $S = 0.60$	$F = 0.86$ $P = 0.93$ SI = 0.53 MH = 0.53	2 weeks; All groups $F = 0.86$ $P = 0.93$ SI = 0.89 MH = 0.85	44	SRS-22 and SF-12; F = 0.21–0.36 $P = 0.29$–0.41 SI = 0.30–0.38 MH = 0.30–0.62 $S = 0.08$–0.44 (AIS and NCSS groups)	$F = 0$ $P = 0$ SI = 0 MH = 0 $S = 1.2$ (AIS patients group)	$F = 0$ $P = 19.4$ SI = 1.1 MH = 8 $S = 9.4$ (AIS patients group)
	Score[b]	−	−	+	−	−		
German								
Niemeyer et al. [24]	Study results	$F = 0.67$ $P = 0.75$ SI = 0.84 MH = 0.88 $S = 0.61$	NA	30 days; $F = 0.80$ $P = 0.76$ SI = 0.87 MH = 0.85 $S = 0.75$	NA	SRS-22 and RMDQ; $F = 0.60$ $P = 0.48$ MH = 0.14	$F = 0$ $P = 0$ SI = 0 MH = 0 $S = 0$	$F = 1.3$ $P = 17.9$ SI = 7.7 MH = 11.5 $S = 26.9$
	Score[b]	−	NA	+	NA	−		
Greek								
Antonarakos et al. [25]	Study results	$F = 0.75$ $P = 0.85$ SI = 0.83 MH = 0.87 $S = 0.67$	NA	30 days; $F = 0.93$ $P = 0.82$ SI = 0.83 MH = 0.79 $S = 0.72$	NA	SRS-22 and SF-36; $F = 0.65$–0.79 $P = 0.61$–0.87 SI = 0.59–0.74 MH = 0.64–0.89 $S = 0.38$–0.63	$F = 2$ $P = 2$ SI = 2 MH = 2 $S = 2$	$F = 7.8$ $P = 9.8$ SI = 13.7 MH = 2 $S = 37.3$
	Score[b]	+	NA	+	NA	+		

Table 4 continued

Language		Internal consistency (Cronbach's α)	Measurement error (SEM)	Reliability (test/retest time intervals and intraclass coefficient correlations by subscale)	Structural Validity (% of explained variance)	Hypotheses Testing[a] (measure comparisons and Pearson/Spearman correlation coefficients)	Floor effects (%)	Ceiling effects (%)
Potoupnis et al. [26]	Study results	$F = 0.73$ $P = 0.83$ SI = 0.89 MH = 0.91 $S = 0.66$	NA	2 weeks; $F = 0.78$ $P = 0.81$ SI = 0.88 MH = 0.82 $S = 0.79$	Estimated, % not reported	SRS-22 and SF-36; $F = 0.39$–0.52 $P = 0.19$–0.37 SI = 0.42–0.63 MH = 0.51–0.75 $S = 0.20$–0.33	$F = 1.1$ $P = 1.1$ SI = 1.1 MH = 1.1 $S = 1.1$	$F = 52.9$ $P = 18.4$ SI = 11.5 MH = 8 $S = 31$
	Score[b]	+	NA	+	?	−		
Italian								
Monticone et al. [27]	Study results	$F = 0.65$ $P = 0.75$ SI = 0.76 MH = 0.78 $S = 0.70$	NA	7 days; $F = 0.99$ $P = 0.98$ SI = 0.96 MH = 0.97 $S = 0.97$	54	SRS-22 and SF-36; $F = 0.32$–0.34 $P = 0.40$–0.65 SI = 0.32–0.42 MH = 0.54–0.79 $S = 0.09$–0.25	NA	$F = 0$ $P = 0$ SI = 0 MH = 0 $S = 0$
	Score[b]	+	NA	+	+	−		
Japanese								
Hashimoto et al. [28]	Study results	$F = 0.65$ $P = 0.76$ SI = 0.74 MH = 0.84	NA	NA	Estimated, % not reported	SRS-22 and SF-36; $F = 0.52$–0.67 $P = 0.44$–0.73 SI = 0.13–0.36 MH = 0.54–0.80	NA	$F = 38$ $P = 36$ SI = 1 MH = 15
	Score[b]	+	NA	NA	?	−		
Korean								
Lee et al. [29]	Study results	$F = 0.85$ $P = 0.83$ SI = 0.75 MH = 0.81 $S = 0.61$	NA	4 weeks; $F = 0.83$ $P = 0.81$ SI = 0.84 MH = 0.88 $S = 0.87$	NA	SRS-22 and SF-36; $F = 0.54$–0.78 $P = 0.71$–0.81 SI = 0.37–0.76 MH = 0.36–0.61 $S = 0.27$–0.44	$F = 1.2$ $P = 1.2$ SI = 2.4 MH = 1.2 $S = 2.4$	$F = 31.3$ $P = 42.1$ SI = 4.8 MH = 12 $S = 8.4$
	Score[b]	+	NA	+	NA	+		
Norwegian								
Adobor et al. [30]	Study results	$F = 0.87$ $P = 0.93$ SI = 0.93 MH = 0.89 $S = 0.90$	NA	2 weeks; $F = 0.76$ $P = 0.87$ SI = 0.87 MH = 0.80 $S = 0.82$	NA	SRS-22 and EuroQol; $F = 0.36$ $P = 0.59$ SI = 0.62 MH = 0.57	$F = 0$ $P = 0$ SI = 1.8 MH = 0 $S = 4.2$	$F = 0$ $P = 10.5$ SI = 0 MH = 10 $S = 7.3$
	Score[b]	+	NA	+	NA	+		
Persian								
Mousavi et al. [31]	Study results	$F = 0.70$ $P = 0.73$ SI = 0.68 MH = 0.78 $S = 0.76$	NA	2 days; $F = 0.87$ $P = 0.82$ SI = 0.85 MH = 0.79 $S = 0.81$	NA	SRS-22 and SF-36; $F = 0.54$–0.67 $P = 0.48$–0.74 SI = 0.45–0.55 MH = 0.66–0.85 $S = 0.35$–0.55	$F = 2.8$ $P = 2.8$ SI = 2.8 MH = 2.8 $S = 2.8$	$F = 2.8$ $P = 16.1$ SI = 2.8 MH = 5.8 $S = 19.4$
	Score[b]	+	NA	+	NA	+		

Table 4 continued

Language		Internal consistency (Cronbach's α)	Measurement error (SEM)	Reliability (test/retest time intervals and intraclass coefficient correlations by subscale)	Structural Validity (% of explained variance)	Hypotheses Testing[a] (measure comparisons and Pearson/Spearman correlation coefficients)	Floor effects (%)	Ceiling effects (%)
Polish								
Glowacki et al. [32]	Study results	$F = 0.81$ $P = 0.81$ SI = 0.77 MH = 0.80 $S = 0.69$	NA	NA	NA	NA	$F = 1.8$ $P = 1.8$ SI = 1.8 MH = 1.8	$F = 10$ $P = 15$ SI = 16.7 MH = 6.7
	Score[b]	+	NA	NA	NA	NA		
Spanish								
Bago et al. [33]	Study results	$F = 0.67$ $P = 0.81$ SI = 0.73 MH = 0.83 $S = 0.78$	NA	1 week; $F = 0.82$ $P = 0.93$ SI = 0.94 MH = 0.94 $S = 0.98$	NA	NA	$F = 0.6$ $P = 0.6$ SI = 1.1 MH = 0.6 $S = 0.6$	$F = 1.1$ $P = 25.7$ SI = 1.7 MH = 10.3 $S = 41.7$
	Score[b]	+	NA	+	NA	NA		
Climent et al. [34]	Study results	NA	NA	NA	56	SRS-22 and Quality of Life for Spine Deformities Profile; $F = 0.52$ $P = 0.85$ SI = 0.62	NA	NA
	Score[b]	NA	NA	NA	+	+		
Swedish								
Danielsson et al. [35]	Study results	$F = 0.72$ $P = 0.78$ SI = 0.84 MH = 0.87 $S = 0.81$	NA	2 weeks; $F = 0.87$ $P = 0.93$ SI = 0.78 MH = 0.80 $S = 0.84$	NA	SRS-22 and SF-36; $F = 0.36–0.56$ $P = 0.45–0.74$ SI = 0.36–0.47 MH = 0.58–0.88 $S = 0.08–0.27$	$F = 0.5$ $P = 0.5$ SI = 0.5 MH = 0.5 $S = 1.7$	$F = 21.8$ $P = 28$ SI = 5.7 MH = 11.9 $S = 17.4$
	Score[b]	+	NA	+	NA	−		
Thai								
Leelapattana et al. [36]	Study results	$F = 0.83$ $P = 0.72$ SI = 0.87 MH = 0.83 $S = 0.63$	NA	10 days; $F = 0.81$ $P = 0.72$ SI = 0.85 MH = 0.82 $S = 0.62$	NA	SRS-22 and SF-36; $F = 0.45–0.62$ $P = 0.45–0.77$ SI = 0.43–0.63 MH = 0.47–0.75 $S = 0.03–0.10$	$F = 6.7$ $P = 3.3$ SI = 3.3 MH = 3.3 $S = 6.7$	$F = 33.3$ $P = 23.3$ SI = 6.7 MH = 3.3 $S = 40$
	Score[b]	+	NA	+	NA	+		
Sathira-Angkura et al. [37]	Study results	$F = 0.70$ $P = 0.76$ SI = 0.81 MH = 0.80 $S = 0.73$	NA	14 days; $F = 0.79$ $P = 0.84$ SI = 0.89 MH = 0.90 $S = 0.84$	NA	SRS-22 and SF-36; $F = 0.45–0.73$ $P = 0.42–0.73$ SI = 0.33–0.49 MH = 0.47–0.68 $S = 0.07–0.35$	F = 1.7 $P = 1.7$ SI = 1.7 MH = 1.7 $S = 1.7$	F = 15.5 $P = 13.8$ SI = 3.4 MH = 6.9 $S = 43.1$
	Score[b]	+	NA	+	NA	−		

Table 4 continued

Language		Internal consistency (Cronbach's α)	Measurement error (SEM)	Reliability (test/retest time intervals and intraclass coefficient correlations by subscale)	Structural Validity (% of explained variance)	Hypotheses Testing[a] (measure comparisons and Pearson/Spearman correlation coefficients)	Floor effects (%)	Ceiling effects (%)
Turkish								
Alanay et al. [38]	Study results	$F = 0.48$ $P = 0.72$ SI = 0.81 MH = 0.72 $S = 0.83$	NA	35 days; $F = 0.76$ $P = 0.63$ SI = 0.82 MH = 0.78 $S = 0.81$	NA	SRS-22 and SF-36; $F = 0.37$–0.63 $P = 0.49$–0.75 SI = 0.34–0.65 MH = 0.68–0.81 $S = 0.27$–0.50	$F = 2.1$ $P = 2.1$ SI = 2.1 MH = 2.1 $S = 4.3$	$F = 2.1$ $P = 17$ SI = 8.5 MH = 6.4 S = 55.3
	Score[b]	+	NA	+	NA	+		

Content validity, criterion validity, and responsiveness are not reported in the table, since no studies included in the review evaluated these measurement properties

NA not available, *SRS-22*, Scoliosis Research Society-22 Patient Questionnaire, *F* Function, *P* Pain, *SI* Self-image, *MH* Mental Health; *S* Satisfaction, *SF-36* Short-Form Health Survey-36 items, *CHQ-CF87* Child Health Questionnaire-Child Form 87, *VAS* visual analogue scale, *SF-12* Short-Form Health Survey-12 items, *RMDQ* Roland and Morris Disability Questionnaire

[a] Hypotheses testing was evaluated based on the correlation coefficients between the relevant domains of SRS-22 and those of the comparative instrument assessing the same construct

[b] Score (+, positive rating; ?, indeterminate rating; −, negative rating) was defined based on the quality criteria for measurement properties proposed by Terwee et al. [12]

Table 5 Assessment of level of evidence (based on the Cochrane Back Review Group [13, 14])

Language study	Internal consistency	Measurement error	Reliability	Structural validity	Hypotheses testing
Brazilian [15, 16]	+	NA	+	NA	−
Chinese (traditional) 17,18]	++	NA	+	+	+
Chinese (simplified) [19, 20]	±	NA	?	NA	±
Dutch [21]	+	NA	+	NA	+
French Canadian [22, 23]	−	−	+	−	−
German [24]	−	NA	?	NA	−
Greek [25, 26]	++	NA	+	?	±
Italian [27]	+++	NA	+	+++	−
Japanese [28]	+	NA	NA	?	+
Korean [29]	+	NA	?	NA	+
Norwegian [30]	+	NA	+	NA	+
Persian [31]	+	NA	?	NA	+
Polish [32]	+	NA	NA	NA	NA
Spanish [33, 34]	+	NA	+	++	+
Swedish [35]	+	NA	?	NA	−
Thai [36, 37]	++	NA	+	NA	±
Turkish [38]	+	NA	?	NA	+

Content validity, criterion validity and responsiveness are not reported in the table, since no studies included in the review evaluated these measurement properties

+++ or −−−, strong evidence positive/negative result; ++ or −−, moderate evidence positive/negative result; + or −, limited evidence positive/negative result; ±, conflicting evidence; ?, unknown, due to poor methodological quality; NA, no information available

NA (not available) means that the psychometric property was not investigated

Brazilian

Two studies were available [15, 16]. The methodological quality of the translation was poor as pretest was not performed [16]. There was limited positive evidence for internal consistency ($\alpha = 0.70$–0.85 and fair methodological quality due to inadequate sample size and unavailable information on missing items [16]). Reliability had limited positive evidence: ICCs ≥ 0.90 and fair methodological quality due to test–retest interval <14 days, as recommended [12], and low sample size [16]. Limited negative evidence was reported for hypotheses testing: It gained a negative result ($r < 0.50$), and the methodological quality was fair because of inadequate sample size and unclear hypotheses [15]. Function, Pain, and Management Satisfaction/Dissatisfaction showed ceiling effects (i.e., >15 % of respondents achieved the highest possible score) [16].

Chinese (traditional)

Two studies were available [17, 18]. The methodological quality of the translation was poor as both did not perform pretesting. There was moderate positive evidence for internal consistency: $\alpha \geq 0.70$ and fair methodological quality due to unavailable information on missing items [17, 18]. Reliability had limited positive evidence: ICC's ≥ 0.76 but poor to fair methodological quality due to inadequate test–retest intervals. Structural validity had limited positive evidence: The explained variance was equal to 67.66 %, and the methodological quality was fair because the percentage of missing items was not reported [17]. Limited positive evidence was reported for hypotheses testing: $r \geq 0.50$ but with an unclear formulation of the hypotheses [18]. Function [18], Pain [17, 18], and Mental Health [18] showed ceiling effects.

Chinese (simplified)

Two studies were available [19, 20]. The methodological quality of the translation was poor as both did not perform pretesting. There was conflicting evidence for internal consistency: $\alpha > 0.70$ in one study [19] and $\alpha < 0.70$ in the other [20]; the methodological quality of both was fair as the percentage of missing items was not reported [19, 20]. Reliability had unknown evidence: It achieved a positive result (ICCs ≥ 0.70), but the methodological quality was poor as test–retest conditions were different [19, 20]. Conflicting evidence was reported on hypotheses testing: It demonstrated a positive result in the former study ($r > 0.50$ [19]) and a negative result in the latter ($r < 0.50$ [20]); the methodological quality was fair as hypotheses were unclearly declared [19, 20]. Function [20], Pain [19, 20], Self-image [20], and Mental Health [20] showed ceiling effects.

Dutch

One study was available [21]. The methodological quality of the translation was excellent. There was limited positive evidence for internal consistency and reliability: α/ICCs ≥ 0.70 with fair methodological quality as percentage of missing items was not reported. Limited positive evidence was reported on hypotheses testing: $r \geq 0.50$, but the hypotheses were unclearly stated. Function, Pain, and Management Satisfaction/Dissatisfaction showed ceiling effects.

French Canadian

Two studies were available [22, 23], but only one provided a translation, which was rated fair as the characteristics of the subjects enrolled for pretesting were inadequately described [22]. There was moderate negative evidence for internal consistency: $\alpha < 0.70$ and fair methodological quality as the percentage of missing items was not reported [22, 23]. Measurement error and reliability had limited negative and positive evidence, respectively: MIC < SDC and ICCs ≥ 0.85; in both cases, the sample size was too small [23]. Moderate negative evidence was reported on structural validity: The explained variance was <50 %, and the methodological quality was fair as the percentage of missing items was not reported [22, 23]. Moderate negative evidence was showed on hypotheses testing: $r < 0.50$ with a fair methodological quality since handling of missing items was not described [22, 23]. Pain [22, 23] and Management Satisfaction/Dissatisfaction [22] showed ceiling effects.

German

One study was available [24]. The methodological quality of the translation was fair as pretest involved a small sample of healthy subjects. Limited negative evidence was reported for internal consistency ($\alpha < 0.70$ and fair methodological quality due to unavailable information about missing items). Reliability had unknown evidence: It showed a positive result (ICCs ≥ 0.75), but the methodological quality was poor as test–retest interval was far too long. Hypotheses testing resulted in limited negative evidence ($r < 0.50$ and unclear formulation of the hypotheses). Pain and Management Satisfaction/Dissatisfaction showed ceiling effects.

Greek

Two studies were available [25, 26]. The methodological quality of the translation was fair in one study (pre-test on healthy subjects [25]) and poor in the other one (absence of

pretest [26]). There was moderate positive evidence for internal consistency: $\alpha \geq 0.70$ and fair methodological quality as the percentage of missing items was not reported [25, 26]. Reliability had limited positive evidence: ICCs > 0.70 with poor and fair methodological quality (small sample size and long test–retest interval in one study [25]; no information on missing items in the other [26]). Structural validity provided unknown evidence: Explained variance was not reported, and the methodological quality was fair as percentage of missing items was not reported [26]. Hypotheses testing showed conflicting evidence: positive in one study ($r > 0.50$ [25]) and negative in the other ($r < 0.50$ [26]); the methodological quality was fair because of unclear hypotheses. Function [26], Pain [26], and Management Satisfaction/Dissatisfaction [25, 26] showed ceiling effects.

Italian

One study was available [27]. The methodological quality of the translation was excellent. There was strong positive evidence for internal consistency: $\alpha \geq 0.70$ with a good methodological quality. Reliability had limited positive evidence: ICCs ≥ 0.96 with a fair methodological quality (short test–retest interval). There was strong positive evidence on structural validity: The explained variance was equal to 54 %, and the methodological quality was excellent. Hypotheses testing demonstrated limited negative evidence: $r < 0.50$ and unclear formulation of hypotheses.

Japanese

One study was available [28]. The methodological quality of the translation was poor (no pretest). There was limited positive evidence for internal consistency: $\alpha \geq 0.70$ and fair methodological quality (percentage of missing not reported). Structural validity displayed unknown evidence. Limited positive evidence was reported on hypotheses testing: $r \geq 0.50$, but unclear formulation of hypotheses. Function, Pain, and Mental Health showed ceiling effects.

Korean

One study was available [29]. The methodological quality of the translation was rated as excellent. There was limited positive evidence for internal consistency: $\alpha \geq 0.70$ with fair methodological quality (percentage of missing items not reported). Unknown evidence was reported for reliability: ICCs ≥ 0.81, but the methodological quality was poor as test–retest interval was too long. Hypotheses testing resulted in limited positive evidence: $r \geq 0.50$, but unclear hypotheses formulation. Function and Pain showed ceiling effects.

Norwegian

One study was available [30]. The methodological quality of the translation was rated as poor because it did not include pretesting. There was limited positive evidence for internal consistency and reliability: α/ICCs ≥ 0.70 with fair methodological quality (percentage of missing items not reported). Hypotheses testing resulted in limited positive evidence: $r \geq 0.50$, but the methodological quality was fair because of unclear hypotheses.

Persian

One study was available [31]. The methodological quality of the translation was rated as fair since the samples of healthy and scoliosis subjects enrolled in the pretest were inadequately described. There was limited positive evidence for internal consistency: $\alpha \geq 0.70$ and fair methodological quality as the study design showed minor methodological flaws. Unknown evidence was reported for reliability: ICCs ≥ 0.79, but the methodological quality was rated poor because test–retest interval was too short. Hypotheses testing resulted in limited positive evidence: $r \geq 0.50$, but unclear hypotheses formulation. Pain and Management Satisfaction/Dissatisfaction showed ceiling effects.

Polish

One study was available [32]. The methodological quality of the translation was poor because it did not include pretesting. Concerning internal consistency, the results were positive ($\alpha \geq 0.70$), but the methodological quality was fair (percentage of missing items was not reported). Pain and Self-image showed ceiling effects.

Spanish

Two studies were available [33, 34]. The methodological quality of the translation was poor as pretest was not performed [33]. There was limited positive evidence for internal consistency: $\alpha \geq 0.70$ with fair methodological quality (percentage of missing items was reported) [33]. As for reliability, the results were positive (ICCs ≥ 0.82), but the methodological quality was fair (short test–retest interval and small sample) [33]. Structural validity showed moderate positive evidence: The explained variance was 56 %, and it met the standards for good methodological quality [34]. Hypotheses testing resulted in limited positive evidence: $r > 0.50$, but the sample was insufficient for reaching a good level of methodological quality [34]. Pain and Management Satisfaction/Dissatisfaction showed ceiling effects [33].

Swedish

One study was available [35]. The methodological quality of the translation was excellent. The internal consistency was rated as positive ($\alpha \geq 0.70$), and the methodological quality was fair, since authors referred to factor analysis performed on a different study population. Reliability had unknown evidence: It showed a positive result (ICCs ≥ 0.78), but the methodological quality was poor as test–retest conditions were different. Hypotheses testing resulted in limited negative evidence: $r < 0.50$, but the methodological quality was fair due to unclear hypotheses. Function, Pain, and Management Satisfaction/Dissatisfaction showed ceiling effects.

Thai

Two studies were available [36, 37]. The methodological quality of translation was poor in one study (absence of pretest [36]) and fair in the other one (pretest on healthy subjects [37]). The internal consistency was judged as positive ($\alpha \geq 0.70$), and the methodological quality was fair (inadequate sample size [36] and percentage of missing items not reported [36, 37]). Concerning reliability, the results were positive (ICCs ≥ 0.70), but the methodological quality ranged from poor to fair (in one study test–retest conditions were different [36] while in the other the sample size was inadequate and the test–retest conditions were unclearly stated [37]). Hypotheses testing reported conflicting evidence: positive in [36] and negative in [37]; in both cases, the methodological quality was fair (unclear hypotheses [36, 37] and inadequate sample size [36]). Function [36, 37], Pain [36], and Management Satisfaction/Dissatisfaction [36, 37] showed ceiling effects.

Turkish

One study was available [38]. The methodological quality of the translation was poor as pretest was not performed. There was limited positive evidence for internal consistency: $\alpha \geq 0.70$ with fair methodological quality (percentage of missing items not reported and small sample size). Reliability had unknown evidence: It gained a positive result (ICCs ≥ 0.70), but the methodological quality was poor (long test–retest interval). Limited positive evidence was reported on hypotheses testing: $r \geq 0.5$, but the methodological quality was fair (unclear hypotheses and inadequate sample size). Pain and Management Satisfaction/Dissatisfaction showed ceiling effects.

Discussion

Our manuscript aimed at assessing the methodological quality of the measurement properties of the available translations of the SRS-22, evaluating their psychometric estimates, and providing the current level of evidence per language. Translated versions of the SRS-22 were evaluated in 17 different languages. In the majority of the cases (10 out of 17 languages), the methodological quality of the translation process was poor and none of the included studies performed a cross-cultural validation. For each translation, at least half of the information concerning measurement properties was lacking and the evidence for the quality of measurement properties was limited due to methodological shortcomings.

The methodological quality of the measurement properties was partially limited by an inadequate sample size [39, 40], missing values not reported [41], inadequate test–retest time intervals, unclear or different administration conditions [42], high ceiling effects [12], and lack of hypotheses testing [42]. The absence of full descriptions of cross-cultural processes prevented us from understanding whether the constructs underlying the original questionnaire were adequately reflected in the translations, something that can affect the performance of the investigated measurement properties [7]. Only few studies performed exploratory or confirmatory factor analysis of the translated scale, while the majority referred to the factor analysis of the original version, which showed a four-factor structure of the 20 non-management domain questions, explaining 98 % of the total variance [43]. However, it has still to be demonstrated whether this subscale distribution of items is confirmed in different social contexts and languages. About one-third of the studies made exploratory analyses, but it is worth underlining the usefulness of confirmatory approaches, which are more appropriate when hypotheses are made about the dimensions of an instrument [39, 40]. Moreover, a lack of hypotheses testing can lead to a high risk of bias when interpreting the results because it is retrospectively tempting to argue in favor of alternative explanations for low correlations instead of concluding that an instrument is not valid; testing hypotheses also makes the validation process more transparent because it makes it easier to quantify the extent of the correlations [42]. Finally, our findings showed that several SRS-22 translations suffered from ceiling effects, potentially leading to an overestimation of reliability and an underestimation of responsiveness [12]. High ceiling effects were found in particular for the Pain subscale, suggesting that SRS-22 might be inadequate to distinguish subjects with scoliosis suffering from pain from those not suffering from pain. This might suggest to combine the use of the SRS-22 with another questionnaire, specifically evaluating pain.

Since the COSMIN checklist was a new and advanced tool that required high standards for methodological quality and most studies included in this SR were published before its development [17–19, 22, 24, 25, 28, 32–34, 38], it is not

surprising that its criteria were hardly fulfilled. As expected, most of the methodological quality of the measurement properties were poor/fair or incomplete and negatively affected the level of evidence. Additional difficulties lay on the fact that most of the languages included in this SR provided a unique study, meaning that the translation available should present high standards of methodological quality in order to obtain a consistent level of evidence. Two other COSMIN-based systematic reviews, which have investigated the psychometric properties of questionnaires concerning spinal disorders [44, 45], found a number of flaws similar to ours. Future studies on psychometric validation of a health-related questionnaire should therefore take into account the standards for good methodological quality proposed by the COSMIN checklist.

The quality assessment for the psychometric properties was affected by a lack of information concerning some measurement properties investigated (measurement error, structural validity, and responsiveness above all). Nevertheless, available data show positive ratings for more than two-thirds of the investigated properties (internal consistency, reliability, and hypotheses testing) indicating reliable and moderate results.

Overall, most of the included studies investigated only few measurement properties, mainly internal consistency, reliability, structural validity, and hypotheses testing. This leaves a lack of evidence regarding crucial properties, such as responsiveness, preventing us from deriving a definitive evaluation of SRS-22 translations from the results of this SR. Based on the available results and on the criteria used to evaluate the level of evidence, the Chinese (traditional), Dutch, Italian, Norwegian, and Spanish translations are advisable for a clinical and research use, with ≥ 3 of the investigated measurement properties rated as positive, although in many cases with a limited level of evidence. Some encouraging results can be derived also from the analysis of the Greek, Japanese, Korean, Persian, Thai, and Turkish translations, with two properties evaluated as positive with a limited or moderate level of evidence and no properties with clear negative results. The Brazilian, Chinese (simplified), Polish, and Swedish translations have shown contradictory or scarce results, with half positive and half negative ratings or only one property being investigated, and therefore, no suggestions can be formulated. Finally, the French Canadian and German translations received a negative rating for most of the investigated properties with a limited or moderate level of evidence, suggesting to use a different questionnaire to assess HRQoL in subjects with scoliosis in these countries. However, these considerations have to be interpreted with caution, and more high-quality and comprehensive studies should be endorsed in order to clarify whether SRS-22 is characterized by good measurement properties while assessing HRQoL across different countries.

This study has some limitations. Firstly, the exclusion of papers written in languages other than English may have introduced a selection bias; however, the most important studies are published in international English-language journals, and the large number of retrieved articles supports our belief that we identified the most important translations. Secondly, the COSMIN checklist has some items that require a subjective judgement, which may lead to disagreements between reviewers; however, we tested the checklist with all of the reviewers before assessing the methodological quality in order to improve the consistency of the ratings. Thirdly, the original version of the SRS-22 as well as other questionnaires specifically created for the evaluation of scoliosis was deliberately not included in this SR, whose objective was to evaluate the measurements properties of the translated versions of the SRS-22. In the future, more comprehensive studies, involving other existing questionnaires investigating HRQoL in subjects with scoliosis, should be carried out in order to promote for each population an evidence-based choice for the questionnaire with the best measurement properties.

Conclusions

Translated versions of SRS-22 have been evaluated in 17 different languages. The methodological quality of the translation process was poor to fair in most cases. Information regarding the measurement properties is still lacking, and available evidence on the measurement properties is mostly limited. Recommendations on the available translations were provided although it is advisable to use the available translated questionnaires cautiously. Confirmatory factor analyses are strongly advised to provide full cross-cultural adaptations of the SRS-22, and further studies on their psychometric properties are still needed to provide definite conclusions on most of the translated versions.

Acknowledgments The authors would like to thank Kevin Smart for his help in preparing the English version of this manuscript.

References

1. Asher, M. A., Lai, S. M., Burton, D., & Manna, B. (2003). The reliability and concurrent validity of the SRS-22 patient questionnaire for idiopathic scoliosis. *Spine, 28*(1), 63–69.
2. Asher, M. A., Lai, S. M., Burton, D., & Manna, B. (2003). Scoliosis Research Society-22 Patient Questionnaire: Responsiveness to change associated with surgical treatment: Preliminary results. *Spine, 28*(1), 70–73.

3. Asher, M. A., Lai, S. M., & Burton, D. C. (2000). Further development and validation of the Scoliosis Research Society (SRS) outcomes instrument. *Spine, 25*(18), 2381–2386.
4. Asher, M. A., Lai, S. M., Burton, D., & Manna, B. (2003). Discrimination validity of the Scoliosis Research Society-22 Patient Questionnaire: Relationship to idiopathic scoliosis curve pattern and curve size. *Spine, 28*(1), 74–78.
5. Asher, M. A., Lai, S. M., Glattes, C., Burton, D. C., Alanay, A., & Bago, J. (2006). Refinement of the SRS-22 health-related quality of life questionnaire function domain. *Spine, 31*(5), 593–597.
6. Lai, S. M., Asher, M. A., & Burton, D. (2006). Estimating SRS-22 quality of life measures with SF-36. Application in idiopathic scoliosis. *Spine, 31*(4), 473–478.
7. de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. J. (2011). *Measurement in medicine: A practical guide*. Cambridge: Cambridge University Press.
8. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research, 19*(4), 539–549.
9. Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., et al. (2010). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Medical Research Methodology, 10*, 22.
10. Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W., Bouter, L. M., & de Vet, H. C. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. *Quality of Life Research, 21*(4), 651–657.
11. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63*(7), 737–745.
12. Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34–42.
13. van Tulder, M., Furlan, A., Bombardier, C., Bouter, L., & Editorial Board CBRG. (2003). Updated method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group. *Spine, 28*(12), 1290–1299.
14. Furlan, A. D., Pennick, V., Bombardier, C., van Tulder, M., & Editorial Board CBRG. (2009). 2009 updated method guidelines for systematic reviews in the Cochrane Back Review Group. *Spine, 34*(18), 1929–1941.
15. Rosanova, G. C., Gabriel, B. S., Camarini, P. M., Gianini, P. E., Coehlo, D. M., & Oliveira, A. S. (2010). Concurrent validity of the Brazilian version of SRS-22r with Br-SF-36. *Revista Brasileira de Fisioterapia, 14*(2), 121–126.
16. Camarini, P. M., Rosanova, G. C., Gabriel, B. S., Gianini, P. E., & Oliveira, A. S. (2013). The Brazilian version of the SRS-22r questionnaire for idiopathic scoliosis. *Brazilian Journal of Physical Therapy, 17*(5), 494–505.
17. Zhao, L., Zhang, Y., Sun, X., Du, Q., & Shang, L. (2007). The Scoliosis Research Society-22 Questionnaire adapted for adolescent idiopathic scoliosis patients in China: Reliability and validity analysis. *Journal of Children's Orthopaedics, 1*(6), 351–355.
18. Cheung, K., Senkoylu, A., Alanay, A., Genc, Y., Lau, S., & Luk, K. D. (2007). Reliability and concurrent validity of the adapted Chinese version of Scoliosis Research Society-22 (SRS-22) Questionnaire. *Spine, 32*(10), 1141–1145.
19. Li, M., Wang, C. F., Gu, S. X., He, S. S., Zhu, X. D., Zhao, Y. C., & Zhang, J. T. (2009). Adapted simplified Chinese (mainland) version of Scoliosis Research Society-22 Questionnaire. *Spine, 34*(12), 1321–1324.
20. Qiu, G., Qiu, Y., Zhu, Z., Liu, Z., Song, Y., Hai, Y., et al. (2011). Re-evaluation of reliability and validity of simplified Chinese version of SRS-22 patient questionnaire: A multicenter study of 333 cases. *Spine, 36*(8), E545–E550.
21. Schlösser, T. P., Stadhouder, A., Schimmel, J. J., Lehr, A. M., van der Heijden, G. J., & Castelein, R. M. (2013). Reliability and validity of the adapted Dutch version of the revised Scoliosis Research Society 22-item Questionnaire. *The Spine Journal,*. doi:10.1016/j.spinee.2013.09.046.
22. Beausejour, M., Joncas, J., Goulet, L., Roy-Beaudry, M., Parent, S., Grimard, G., et al. (2009). Reliability and validity of adapted French Canadian version of Scoliosis Research Society outcomes Questionnaire (SRS-22) in Quebec. *Spine, 34*(6), 623–628.
23. Lonjon, G., Ilharreborde, B., Odent, T., Moreau, S., Glorion, C., & Mazda, K. (2014). Reliability and validity of the French-Canadian version of the Scoliosis Research Society 22 Questionnaire in France. *Spine, 39*(1), E26–E34.
24. Niemeyer, T., Schubert, C., Halm, H. F., Herberts, T., Leichtle, C., & Gesicki, M. (2009). Validity and reliability of an adapted German version of Scoliosis Research Society-22 Questionnaire. *Spine, 34*(8), 818–821.
25. Antonarakos, P. D., Katrinitsa, L., Angelis, L., Paganas, A., Koen, E. M., Christodoulou, E. A., & Christodoulou, A. G. (2009). Reliability and validity of the adapted Greek version of Scoliosis Research Society-22 (SRS-22) Questionnaire. *Scoliosis, 4*, 14.
26. Potoupnis, M., Papavasiliou, K., Kenanidis, E., Pellios, S., Kapetanou, A., Sayegh, F., & Kapetanos, G. (2012). Reliability and concurrent validity of the adapted Greek version of the Scoliosis Research Society-22r Questionnaire. A cross-sectional study performed on conservatively treated patients. *Hippokratia, 16*(3), 225–229.
27. Monticone, M., Baiardi, P., Calabrò, D., Calabrò, F., & Foti, C. (2010). Development of the Italian version of the revised Scoliosis Research Society-22 Patient Questionnaire, SRS-22r-I: Cross-cultural adaptation, factor analysis, reliability, and validity. *Spine, 35*(24), E1412–E1417.
28. Hashimoto, H., Sase, T., Arai, Y., Maruyama, T., Isobe, K., & Shouno, Y. (2007). Validation of a Japanese version of the Scoliosis Research Society-22 Patient Questionnaire among idiopathic scoliosis patients in Japan. *Spine, 32*(4), E141–E146.
29. Lee, J. S., Lee, D. H., Suh, K. T., Kim, J. I., Lim, J. M., & Goh, T. S. (2011). Validation of the Korean version of the Scoliosis Research Society-22 Questionnaire. *European Spine Journal, 20*(10), 1751–1756.
30. Adobor, R. D., Rimeslåtten, S., Keller, A., & Brox, J. I. (2010). Repeatability, reliability, and concurrent validity of the Scoliosis Research Society-22 Questionnaire and EuroQol in patients with adolescent idiopathic scoliosis. *Spine, 35*(2), 206–209.
31. Mousavi, S. J., Mobini, B., Mehdian, H., Akbarnia, B., Bouzari, B., Askary-Ashtiani, A., et al. (2010). Reliability and validity of the Persian version of the Scoliosis Research Society-22r Questionnaire. *Spine, 35*(7), 784–789.
32. Glowacki, M., Misterska, E., Laurentowska, M., & Mankowski, P. (2009). Polish adaptation of Scoliosis Research Society-22 Questionnaire. *Spine, 34*(10), 1060–1065.
33. Bago, J., Climent, J. M., Ey, A., Perez-Grueso, F. J., & Izquierdo, E. (2004). The Spanish version of the SRS-22 patient questionnaire for idiopathic scoliosis: Transcultural adaptation and reliability analysis. *Spine, 29*(15), 1676–1680.
34. Climent, J. M., Bago, J., Ey, A., Perez-Grueso, F. J., & Izquierdo, E. (2005). Validity of the Spanish version of the Scoliosis Research Society-22 Patient Questionnaire. *Spine, 30*(6), 705–709.
35. Danielsson, A. J., & Romberg, K. (2013). Reliability and validity of the Swedish version of the Scoliosis Research Society-22

(SRS-22r) Patient Questionnaire for idiopathic scoliosis. *Spine, 38*(21), 1875–1884.
36. Leelapattana, P., Keorochana, G., Johnson, J., Wajanavisit, W., & Laohacharoensombat, W. (2011). Reliability and validity of an adapted Thai version of the Scoliosis Research Society-22 Questionnaire. *Journal of Children's Orthopaedics, 5*(1), 35–40.
37. Sathira-Angkura, V., Pithankuakul, K., Sakulpipatana, S., Piyaskulkaew, C., & Kunakornsawat, S. (2012). Validity and reliability of an adapted Thai Version of Scoliosis Research Society-22 Questionnaire for adolescent idiopathic scoliosis. *Spine, 37*(9), 783–787.
38. Alanay, A., Cil, A., Berk, H., Acaroglu, R. E., Yazici, M., Akcali, O., et al. (2005). Reliability and validity of adapted Turkish version of Scoliosis Research Society-22 (SRS-22) Questionnaire. *Spine, 30*(21), 2464–2468.
39. Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 286–299.
40. De Vet, H. C. W., Ader, H. J., Terwee, C. B., & Pouwer, F. (2005). Are factor analytical techniques appropriately used in the validation of health status questionnaires? A systematic review on the quality of factor analyses of the SF-36. *Quality of Life Research, 14*(5), 1203–1218.
41. Fayers, P. M., Curran, D., & Machin, D. (1998). Incomplete quality of life data in randomized trials: Missing items. *Statics in Medicine, 17*(5–7), 679–696.
42. Terwee, C. B., Schellingerhout, J. M., Verhagen, A. P., Koes, B. W., & de Vet, H. C. (2011). Methodological quality of studies on the measurement properties of neck pain and disability questionnaires: A systematic review. *Journal of Manipulative and Physiological Therapeutics, 34*(4), 261–272.
43. Lai, S. M., Asher, M. A., Burton, D. C., & Carlson, B. B. (2010). Identification of Scoliosis Research Society-22r health-related quality of life questionnaire domains using factor analysis methodology. *Spine, 35*(12), 1236–1240.
44. Schellingerhout, J. M., Verhagen, A. P., Heymans, M. W., Koes, B. W., de Vet, H. C., & Terwee, C. B. (2012). Measurement properties of disease-specific questionnaires in patients with neck pain: A systematic review. *Quality of Life Research, 21*(4), 659–670.
45. Schellingerhout, J. M., Heymans, M. W., Verhagen, A. P., de Vet, H. C., Koes, B. W., & Terwee, C. B. (2011). Measurement properties of translated versions of neck-specific questionnaires: A systematic review. *BMC Medical Research Methodology, 11*, 87.