Research Report

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

Alessandro Chiarotto, Lara J. Maxwell, Caroline B. Terwee, George A. Wells, Peter Tugwell, Raymond W. Ostelo

A. Chiarotto, MSc, Department of Health Sciences, Faculty of Earth and Life Sciences, EMGO+ Institute for Health and Care Research, VU University, De Boelelaan 1085, Room U-601, 1081HV, Amsterdam, the Netherlands, and Department of Epidemiology and Biostatistics, EMGO+ Institute for Health and Care Research, VU University Medical Center, Amsterdam, the Netherlands. Address all correspondence to Mr Chiarotto at: a.chiarotto@vu.nl.

L.J. Maxwell, MSc, Centre for Practice-Changing Research, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada.

C.B. Terwee, PhD, Department of Epidemiology and Biostatistics, EMGO+ Institute for Health and Care Research, VU University Medical Center.

G.A. Wells, PhD, Department of Epidemiology and Community Medicine, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada.

P. Tugwell, MD, MSc, Department of Medicine, Ottawa Hospital Research Institute, University of Ottawa.

R.W. Ostelo, PhD, Department of Health Sciences, Faculty of Earth and Life Sciences, EMGO+ Institute for Health and Care Research, VU University, and Department of Epidemiology and Biostatistics, EMGO+ Institute for Health and Care Research, VU University Medical Center.

[Chiarotto A, Maxwell LJ, Terwee CB, et al. 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*. 2016;96:1620–1637.]

Published Ahead of Print: April 14, 2016
Accepted: March 31, 2016
Submitted: August 17, 2015

Background. Physical functioning is a core outcome domain to be measured in nonspecific low back pain (NSLBP). A panel of experts recommended the Roland-Morris Disability Questionnaire (RMDQ) and Oswestry Disability Index (ODI) to measure this domain. The original 24-item RMDQ and ODI 2.1a are recommended by their developers.

Purpose. The purpose of this study was to evaluate whether the 24-item RMDQ or the ODI 2.1a has better measurement properties than the other to measure physical functioning in adult patients with NSLBP.

Data Sources. Bibliographic databases (MEDLINE, Embase, CINAHL, SportDiscus, PsycINFO, and Google Scholar), references of existing reviews, and citation tracking were the data sources.

Study Selection. Two reviewers selected studies performing a head-to-head comparison of measurement properties (reliability, validity, and responsiveness) of the 2 questionnaires. The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist was used to assess the methodological quality of these studies.

Data Extraction. The studies' characteristics and results were extracted by 2 reviewers. A meta-analysis was conducted when there was sufficient clinical and methodological homogeneity among studies.

Data Synthesis. Nine articles were included, for a total of 11 studies assessing 5 measurement properties. All studies were classified as having poor or fair methodological quality. The ODI displayed better test-retest reliability and smaller measurement error, whereas the RMDQ presented better construct validity as a measure of physical functioning. There was conflicting evidence for both instruments regarding responsiveness and inconclusive evidence for internal consistency.

Limitations. The results of this review are not generalizable to all available versions of these questionnaires or to patients with specific causes for their LBP.

Conclusions. Based on existing head-to-head comparison studies, there are no strong reasons to prefer 1 of these 2 instruments to measure physical functioning in patients with NSLBP, but studies of higher quality are needed to confirm this conclusion. Foremost, content, structural, and cross-cultural validity of these questionnaires in patients with NSLBP should be assessed and compared.

Low back pain (LBP) is the primary worldwide cause of years lived with disability according to a report of the Global Burden of Disease.[1] Approximately 80% of people experience activity-limiting LBP at some point in their lifetime, and approximately 5% develop chronic LBP lasting for more than 3 months.[2] Costs associated with LBP represent a serious burden to society, and lost work productivity accounts for the bulk of these costs.[3,4] Approximately 90% of patients with LBP are labeled as having nonspecific low back pain (NSLBP) because a specific cause for their pain cannot be found.[5–7]

Limitations in physical functioning are frequently reported by patients with NSLBP. The measurement of physical functioning as a core outcome domain in all clinical trials for NSLBP has been recently recommended by a wide, international, multidisciplinary, and multistakeholder panel of experts.[8] Several patient-reported and back-specific questionnaires have been developed and used to measure back-specific functional status.[9] Among these questionnaires, 2 are most frequently used[10] and were previously recommended by panels of experts[11,12]: the Roland-Morris Disability Questionnaire (RMDQ) and the Oswestry Disability Index (ODI). Different versions of both questionnaires have been developed over time,[9] and to reduce inconsistency across studies, one specific version for each questionnaire was recommended by their developers: the original 24-item RMDQ and version 2.1a of the ODI.[13]

The original RMDQ was developed in 1983 from the Sickness Impact Profile, with the aim of developing "a simple, sensitive, and reliable method of measuring disability in patients with back pain."[14(p141)] It consists of 24 items representing "physical functions that were likely to be affected by LBP"; each item can be checked if it applies to a patient for that day, leading to a total score that is obtained by counting the number of checked items.[13(p3115)] The original version of the ODI (ie, ODI 1.0) was published in 1980 with the scope of being "a valid indicator of disability," where *disability* was defined as "the limitations of a patient's performance compared with that of a fit person."[15(p271)] The ODI consists of 10 items representing different health constructs (eg, pain intensity, physical functioning, sleep functioning, social functioning).[16] The first item of ODI 1.0 underwent a substantial change that resulted in the development of ODI version 2.0,[17] which presented some very small typographical errors that were corrected to become version 2.1a of the questionnaire.[18] The total score of the ODI is calculated by adding all scores of applicable items, dividing the obtained score by the maximal total score, and by multiplying the result by 100 to obtain a percentage score.[16]

To be used in research and clinical practice, a measurement instrument needs to show adequate measurement properties (ie, validity, reliability, and responsiveness).[19] The measurement properties of an instrument are context-specific (ie, they depend on various factors, such as study population, clinical setting, time points of assessment, and comparator instruments).[20] Therefore, to make an adequate judgment on which of 2 instruments has better measurement properties, both instruments should be administered to the same patients, in the same setting, at the same time points, and with the same comparator instruments. For researchers, clinicians, and their patients who want or have to make a choice between recommended versions of RMDQ and ODI, it would be crucial to know whether one instrument has better measurement properties than the other.

An attempt to compare the measurement properties of the RMDQ and ODI has been made in some reviews[13,21–24]; however, all of these reviews failed on some key methodological aspects for systematic reviews on measurement properties of instruments.[20] Two of these reviews were narrative reviews, as they were not conducted in a systematic fashion,[13,24] and none of them included an assessment of the methodological quality of the studies, which was necessary to weight the trustworthiness of results. Moreover, none of them aimed specifically at focusing on head-to-head comparison studies, which have the best design to establish whether an instrument is better than another.[25] Newman et al[26] recently performed a systematic review of head-to-head comparisons between RMDQ and ODI, but they focused only on responsiveness, without making a specific distinction between different versions of the questionnaires, and included all LBP disorders in their evidence synthesis. Hence, to date, no systematic reviews have been conducted to summarize head-to-head comparison studies focusing on all measurement properties of recommended versions of RMDQ and ODI in only patients with NSLBP.

This systematic review purported to determine whether the 24-item RMDQ or the ODI 2.1a has better measurement properties than the other to measure physical functioning in patients with NSLBP. The rationale for focusing this review solely on patients with NSLBP is related to the scope of the ongoing international effort aimed at developing a core outcome set of domains and measurement instruments to be used and reported in all clinical trials conducted in this large subgroup of patients with LBP.[8,27] The highest consensus was reached on the measurement of physical functioning,[8] and as a previous panel of experts suggested both the 24-item RMDQ and the ODI 2.1a for this domain,[11,12] it is essential to assess whether 1 of the 2 instruments has better measurement properties in the NSLBP population.

Method

This review was conducted and reported following the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement.[28,29] A protocol was written a priori and can be accessed on the international prospective register of systematic reviews (http://www.crd.york.ac.uk/PROSPERO/, registration number: CRD42014014803).

Data Sources and Searches

The following biomedical databases were last searched on February 2, 2015, to retrieve eligible articles: MEDLINE (through the interface PubMed), Embase (Embase.com), CINAHL (EBSCOhost), PsycINFO (EBSCOhost), and SportDiscus (EBSCOhost). The search strategy con-

sisted of 3 groups of search terms representing the following components of the research aim: (1) RMDQ and ODI, (2) NSLBP, and (3) measurement properties. The 3 groups of search terms were combined with each other with the Boolean operator "AND," and index and/or title/abstracts terms within each group were combined with the operator "OR." A specific search filter was used for retrieval of studies on measurement properties of instruments in the MEDLINE database.[30] The full electronic search strategies for all databases are presented in eAppendix 1 (available at ptjournal.apta.org). No restrictions of language and time were applied to the search strategies. Google Scholar also was searched twice using the extensive names of the 2 questionnaires; the first 100 hits of each search were last checked on February 12, 2015, for inclusion. References of studies included in other systematic reviews[21–23,26] also were screened. Backward citation tracking was performed by checking the references of the studies deemed as eligible; forward citation tracking was performed in the database Web of Science by screening titles of articles that cited the eligible studies.

Study Selection

A study was included if it met the following criteria: (1) full-text original article (eg, not an abstract, editorial, or review), (2) purpose to evaluate one or more measurement properties of both 24-item RMDQ and ODI 2.1a, and (3) study population of adult patients (ie, >18 years old) with NSLBP. For the scope of this review, considering the very small adjustments in wording highlighted by its developer,[18] the 3 versions of the ODI (ie, 2.0, 2.1, and 2.1a) were included, assessed, and renamed as the same questionnaire (ie, ODI 2.1a). Studies including patients with specific mechanical diagnoses (eg, spinal stenosis, herniated disk) were not included. Studies including patients with the following specific nonmechanical causes for their LBP (eg, infection, cancer, rheumatoid arthritis, ankylosing spondylitis, other inflammatory disorders) also were excluded. Studies including a "mixed" population of patients with LBP were included only if at least 75% of the patients met the inclusion criterion, and the same rule was followed for studies including patients with spinal pain at different levels.

Eligibility criteria were applied independently by 2 reviewers (A.C., L.J.M.) to titles and abstracts of all articles retrieved with literature searches. Full texts of potentially eligible articles were downloaded and assessed against the inclusion criteria by the same 2 reviewers independently. Agreement regarding inclusion was sought in a consensus meeting between reviewers, and in case of disagreements, a third reviewer (R.W.O.) made decisions. If it was not clear which version of the RMDQ or ODI was used in a study, the authors of that study were contacted by email to request this information. The corresponding author of a study was contacted first, and if no answer was received, other authors with a retrievable email address were contacted. If an answer was not received by any of the authors or if the authors were not able to say which version was used, the study was not included. Citation tracking and checking references of other reviews were conducted by 1 reviewer (A.C.) and, when potentially eligible studies were retrieved, their eligibility was screened by 2 reviewers independently (A.C., L.J.M.).

Data Extraction and Quality Assessment

The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist[31,32] was used to assess the methodological quality of the studies. This checklist consists of 9 boxes, each representing a measurement property included in the COSMIN taxonomy: internal consistency, reliability, measurement error, content validity, structural validity, construct validity (hypotheses testing), cross-cultural validity, criterion validity, and responsiveness.[19] Each box contains several items that can each be scored on a 4-point rating scale (ie, poor, fair, good, or excellent). An overall score for the methodological quality of each measurement property for each study is determined by taking the lowest rating of any of the items in a box.[32] The COSMIN consensus-based definitions of measurement properties[19] were used to decide which properties were assessed in a study and which corresponding boxes had to be completed, regardless of the terminology used in the included studies. Assessment of the methodological quality was performed by 2 reviewers independently (A.C., L.J.M.), and in case of disagreements, a third reviewer (R.W.O.) made final decisions.

A customized data extraction form was developed for this review, and extracted data were subsequently reported in tables. The following information was extracted from each included study by one reviewer (A.C.) and double checked by a second reviewer (L.J.M.): characteristics of the studies (ie, country, language, design, clinical setting, inclusion and exclusion criteria, type of intervention, methods for selection of patients, measurement properties assessed, time points of assessment), characteristics of the patients included in the studies (ie, sample size, age, sex, disease characteristics, and RMDQ and ODI scores at baseline), and results on the assessed measurement properties.

Data Synthesis and Analysis

Meta-analysis of different parameters (eg, Cronbach alpha, intraclass correlation coefficient [ICC], Pearson correlation) was conducted for studies assessing the same measurement properties of the 2 questionnaires. Data extracted on characteristics of studies and participants were used to assess whether there was sufficient clinical and methodological homogeneity. Results of different studies were statistically pooled when: (1) participants displayed similar characteristics in terms of age, sex, and RMDQ and ODI baseline scores; (2) participants were assessed with the same time interval; and (3) the same statistical parameters (ie, same statistical models or formulas) were used. Pooled correlation coefficients with their 95% confidence intervals (95% CIs) were calculated using a Fisher z transformation of the correlations.[33] In light of expected between-study error, the DerSimonian and Laird random-effects model was used in the meta-analysis.[34] Statistical heterogeneity of results was assessed using the Q statistic and the I^2. The Q statistic reflects the total amount of variance in the meta-analysis, and the I^2 indexes the propor-

tion of variance that is due to between-study differences and is not sensitive to the number of studies considered.[35] The I^2 values range from 0% to 100%, and values >50% are suggested to represent substantial heterogeneity.[35] Sensitivity analyses excluding studies of poor methodological quality were performed to assess whether the pooled estimates were strongly influenced by the results of these studies. All meta-analyses were performed using the Comprehensive Meta-Analysis 2.1 software (Biostat, Englewood, New Jersey).

The overall rating for a measurement property of each instrument was considered "positive," "indeterminate," or "negative," following adapted international quality criteria for good measurement properties (eAppendix 2, available at ptjournal.apta.org).[36] The criteria for measurement error were modified a priori (eAppendix 2) to enable a straightforward interpretation of results on this property. This interpretation of results would not have been possible if using the original criteria, which take for granted that a study would report parameters of measurement error together with the minimal important change (MIC),[36] although this is often not the case. As suggested by the COSMIN initiative,[20] a best evidence synthesis was performed for each measurement property, taking into account the results, their consistency, and the methodological quality of the studies (eAppendix 3, available at ptjournal.apta.org). One instrument was considered to be better than the other on a given measurement property when it displayed at least a moderate level of evidence with consistent and positive ratings and the other instrument displayed conflicting findings or negative ratings (eAppendix 3). When, for a certain measurement property, both instruments displayed the same level of evidence with consistent and positive ratings, 1 of the 2 instruments was considered better than the other if showing consistently better results in all of the studies. Results for each measurement property were carefully inspected to assess whether a clear difference between instruments could be found in patients with acute or subacute/chronic NSLBP duration.

Role of the Funding Source

The authors acknowledge the Wetenshcappelijk College Fysiotherapie (WCF) of the Royal Dutch Society for Physical Therapy (KNGF) for providing funding for this study. This funding body did not have any role in design, conduct, analysis, or interpretation of data, nor in writing the manuscript and deciding to submit the manuscript for publication. The views expressed here are those of the authors and do not necessarily reflect those of their funding bodies.

Results

Figure 1 presents the flowchart for the study selection process. Nine articles[37–45] were considered as eligible, including a total of 11 studies comparing the measurement properties of the 24-item RMDQ and ODI 2.1a in patients with NSLBP. Eight articles in which a head-to-head comparison of RMDQ and ODI was performed were not included because the recommended versions of the RMDQ and ODI (Fig. 1) were not used: 6 studies[46–51] used the ODI 1.0, 1 study[52] used the "chiropractic version" of the ODI, and 1 study[53] used the 23-item version of the RMDQ. Two articles[54,55] were excluded because they presented the direct comparison of the 2 instruments in patients with specific LBP (Fig. 1). Citation tracking of eligible articles did not add any study to those retrieved through databases and searches of other sources. Table 1 presents the characteristics of the studies and the included participants.

Two studies[40,45] evaluated internal consistency, 4 studies[38,40,43,45] evaluated test-retest reliability, 4 studies[38,40,43,44] evaluated measurement error, 5 studies[40,44,45,54] evaluated construct validity, and 7 studies[37–39,41,43,44] evaluated responsiveness. None of the studies made a direct comparison of the following measurement properties: content validity, structural validity, cross-cultural validity, and criterion validity.[19] Five studies[37,40–43] were conducted only in patients with chronic NSLBP, where chronic NSLBP was defined as the presence of nonspecific LBP for more than 3 months (Tab. 1). Two studies[40,41] were conducted in patients with NSLBP for less than 3 weeks, 2 studies[39,44] included patients with subacute and chronic NSLBP (ie, pain for more than 6 weeks), and 2 studies[38,45] included the whole spectrum of NSLBP duration (Tab. 1). Results on the measurement properties are subdivided and presented in the 3 COSMIN macro domains: reliability, validity, and responsiveness (Tabs. 2 and 3, Fig. 2; eTable, available at ptjournal.apta.org). Eight of the studies included in this review[37,40–42,44,45] assessed the measurement properties of 5 translated and cross-culturally adapted versions of RMDQ and ODI (ie, Brazilian, Norwegian, German, Italian, and Persian). These studies were considered and assessed together with those evaluating the measurement properties of the original versions, as no modifications were made in the structure of the questionnaires (eg, number of items, type of response options) during the process of translation and adaptation.

Internal Consistency

Two studies of poor methodological quality assessed internal consistency.[40,45] These studies were classified as being of poor quality because they calculated the Cronbach alpha of the total scores without assessing or providing evidence that the questionnaires were unidimensional (Tab. 2). Considering this limitation, statistical pooling was not performed, and it remains unknown whether one of the questionnaires has better internal consistency (Tab. 4).

Reliability

Four studies assessed the test-retest reliability of the 2 questionnaires: 2 studies[40,43] were classified as being of poor methodological quality, and the other 2 studies[38,45] were classified as being of fair quality because of the small sample sizes included (Tab. 2). The study by Mousavi et al[45] also was classified as being of fair quality because a short time interval was adopted for the reassessment of the participants (Tab. 2). Statistical pooling was not performed due to discrepancies in time points of assessment and differences in (ICC parameters (Tab. 2). Two studies[38,45] included patients with acute and chronic NSLBP but did not report descriptive statistics on pain duration (Tab. 1), making it impossible to identify different findings

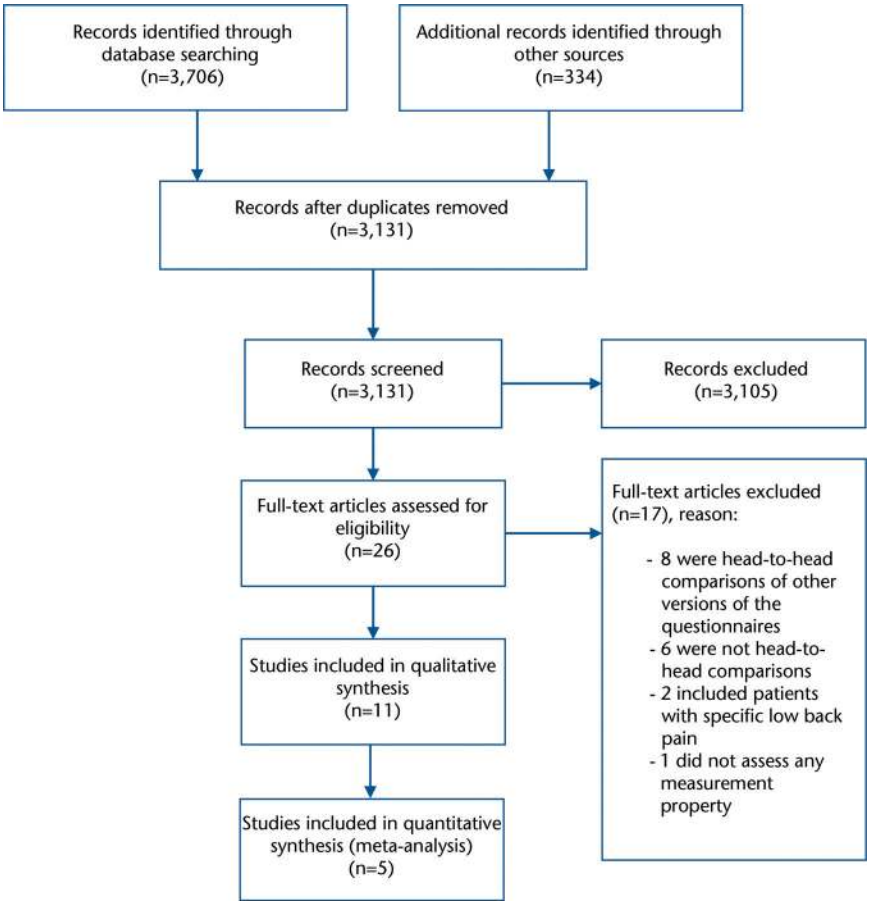

Figure 1.
Flowchart of results of search strategy and selection of articles.

related to NSLBP duration. A moderate level of evidence of good reliability was found for the ODI but not for the RMDQ, for which there were conflicting findings in the 2 studies of fair quality (Tab. 2). These results suggest that the ODI displays better test-retest reliability than the RMDQ (Tab. 4).

Measurement Error

The measurement error of the RMDQ and ODI was compared by 4 studies: 2 studies[40,43] were rated as being of poor methodological quality, and 2 studies[38,44] were rated as being of fair quality. The sample sizes influenced the quality of 3 of these studies (eTable), and the rating of the study by Monticone et al[44] was due to the lack of information on how missing items were handled. Meta-analysis for the standard error of measurement (SEM) and the smallest detectable change (SDC) was not performed due to discrepancies in time intervals and in the parameters' formulas (Tab. 2). Due to limited reporting on NSLBP duration in 3 of these studies,[38,40,44] it was not feasible to identify whether there were discrepant results on this property related to pain duration. In the 2 studies of fair methodological quality, the ODI displayed moderate evidence of a positive rating for its SDC, while the RMDQ displayed a negative rating (Tab. 2; eAppendix 3). These results indicate that the ODI has a smaller measurement error than the RMDQ.

Construct Validity–Hypotheses Testing

Construct validity was assessed in 5 studies: 3 of fair methodological quality[40,45] and 2 of poor methodological quality[42,44] (eTable). Studies of fair quality were judged as such because of limited information regarding the measurement properties of comparator instruments in any study population. The other 2 studies were rated as poor because of lack of information on the comparator instruments[42] or because it was unclear what was expected for the correlations between instruments.[44] Meta-analyses were performed on Pearson correlations of the RMDQ and ODI separately when these correlations were calculated with the same comparator instruments in at least 3 studies (Fig. 2). Given the focus of this review on physical functioning, disease-specific comparator instruments measuring function or disability were considered as instruments measuring the same or a related construct; all other instruments were considered as measuring unrelated constructs.

Pooled correlations with the physical functioning subscale of the Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36-PF) were −.66 (95% CI=−.77, −.60) for the RMDQ and −0.70 (95% CI=−0.77, −0.61) for the ODI (Figs. 2A and 2B); substantial heterogeneity was found for the pooled estimate of the ODI. Pooled correlations of .46 (95% CI=.35, .55) and −.46 (95% CI=−.61, −.26) were found for the RMDQ with pain instruments (Figs. 2C and 2E), and correlations of .54 (95% CI=.41, .64) and −.56 (95% CI=−.68, −.40) were found for the ODI with the same instruments (Figs. 2D and 2F); substantial heterogeneity was found for all but one of these estimates (Fig. 2C). Both instruments displayed correlations below .5 with other unrelated constructs, with the ODI showing higher correlations than the RMDQ and with no substantial heterogeneity in these meta-analyses (Figs. 2G–2J). Sensitivity analyses revealed that all of these pooled estimates were not substantially different when the studies of poor methodological quality[42,44] were removed. Correlations investigated only in 1 or 2 studies were not included in meta-analyses and are presented in the eTable. The ODI showed consistently higher correlations than the RMDQ with all of the other instruments assessed, with the only exception of the correlation with the role–physical subscale of the SF-36 in one study[44] (Fig. 2, eTable). One study[40]

Table 1.
Characteristics of the 11 Studies and of the Patients Included in This Review[a]

Study	Country (Language)	Setting	Study Design	Health Intervention	Study Population	Measurement Properties	n	Patients' Characteristics
Coelho et al,[37] 2008	Brazil (Brazilian Portuguese)	Physical therapy centers	Observational, longitudinal	Physical therapy (including manual therapy, exercises, physical conditioning, and electrotherapy)	NSLBP >3 mo (unspecified whether with or without leg pain)	Responsiveness	30	Sex=33% female Age, $\bar{X}$ (SD)=38.1 (14.1) y PD, $\bar{X}$ (SD)=3.4 (2.5) y PL=not reported WS=47% working PM=40% taking RMDQ, $\bar{X}$ (SD)=11.1 (5.7) ODI, $\bar{X}$ (SD)=32.9 (18.9)
Davidson and Keating,[38] 2002	Australia (English)	Physical therapy centers (3 hospitals, 3 community health services, and 4 private practices)	Observational, longitudinal	Physical therapy	NSLBP with or without leg pain	Reliability Measurement error	47	Sex=64% female Age, $\bar{X}$ (SD)=55.0 (17.0) y PD=not reported PL=17% low back only WS=30% working PM=not reported RMDQ, $\bar{X}$ (SD)=9.0 (5.2) ODI, $\bar{X}$ (SD)=35.0 (15.0)
						Responsiveness	106	Sex=73% female Age, $\bar{X}$ (SD)=49.0 (16.0) y PD=NR PL=39% low back only WS=46% working PM=NR RMDQ, $\bar{X}$ (SD)=9.5 (5.9) ODI, $\bar{X}$ (SD)=35.0 (17.0)
Frost et al,[39] 2008	United Kingdom (English)	Physical therapy centers	Randomized controlled trial	Physical therapy plus one advice session or advice session alone	NSLBP with or without leg pain or neurological signs >6 wk	Responsiveness	201	Sex=55% female Age, $\bar{X}$ (SD)=42.5 (14.4) y PD=NR PL=NR WS=NR PM=NR RMDQ=NR ODI=NR
Grotle et al,[40] 2003[b]	Norway (Norwegian)	Primary health care (GP or chiropractor)	Observational, cross-sectional	NA	NSLBP <2 wk (unspecified whether with or without leg pain)	Construct validity	55	Sex=73% female Age, $\bar{X}$ (SD)=38.0 (10.0) y PD, $\bar{X}$ (SD)=9.5 (7.0) d PL=57% low back only WS=60% working PM=42% taking RMDQ, $\bar{X}$ (SD)=9.0 (5.0) ODI, $\bar{X}$ (SD)=28.0 (15.0)

(Continued)

Table 1.
Continued

Study	Country (Language)	Setting	Study Design	Health Intervention	Study Population	Measurement Properties	n	Patients' Characteristics
Grotle et al,[40] 2003[b]	Norway (Norwegian)	Back clinic of a hospital	Observational, cross-sectional	NA	NSLBP >3 mo (unspecified whether with or without leg pain)	Internal consistency Reliability Measurement error	28	Sex=NR Age=NR PD=NR PL=NR WS=NR PM=NR RMDQ, $\bar{X}$ (SD)=9.8 (4.1) ODI, $\bar{X}$ (SD)=32.0 (12.0)
						Construct validity	50	Sex=62% female Age, $\bar{X}$ (SD)=40.0 (9.0) y PD, $\bar{X}$ (SD)=1.6 (2.2) y PL=12% low back only WS=16% working PM=40% taking RMDQ, $\bar{X}$ (SD)=10.0 (4.0) ODI, $\bar{X}$ (SD)=32.0 (11.0)
Grotle et al,[41] 2004[c]	Norway (Norwegian)	Primary health care (GP or chiropractor)	Observational, longitudinal	Information according to guidelines for acute LBP	NSLBP <3 wk (unspecified whether with or without leg pain)	Responsiveness	52	Sex=73% female Age, $\bar{X}$ (SD)=38.0 (10.0) y PD, $\bar{X}$ (SD)=10.0 (7.0) d PL=57% low back only WS=60% working PM=43% taking RMDQ, $\bar{X}$ (SD)=8.9 (5.1) ODI, $\bar{X}$ (SD)=28.1 (15.4)
Grotle et al,[41] 2004[c]	Norway (Norwegian)	Back clinic of a hospital	Observational, longitudinal	Treatment according to Indahl model for chronic LBP	NSLBP >3 mo (unspecified whether with or without leg pain)	Responsiveness	48	Sex=62% female Age, $\bar{X}$ (SD)=40.0 (9.0) y PD, $\bar{X}$ (SD)=1.6 (2.2) y PL=12% low back only WS=16% working PM=40% taking RMDQ, $\bar{X}$ (SD)=9.5 (4.3) ODI, $\bar{X}$ (SD)=31.8 (11.2)
Mannion et al,[42] 2006	Switzerland (German)	Spine unit of an orthopedic hospital	Observational, cross-sectional	NA	NSLBP >3 mo (unspecified whether with or without leg pain)	Construct validity	32	Sex=59% female Age, $\bar{X}$ (SD)=49.0 (17.0) y PD=NR PL=NR WS=NR PM=NR RMDQ, $\bar{X}$ (SD)=10.1 (5.5) ODI, $\bar{X}$ (SD)=30.5 (17.0)

(Continued)

Table 1.
Continued

Study	Country (Language)	Setting	Study Design	Health Intervention	Study Population	Measurement Properties	n	Patients' Characteristics
Maughan and Lewis,[43] 2010	United Kingdom (English)	Physical therapy back classes	Observational, longitudinal	Physical therapy (including exercises and education)	NSLBP with or without leg pain >3 mo	Reliability Measurement error	25	Sex=60% female Age, $\overline{X}$ (SD)=50.0 (NR) y PD, $\overline{X}$ (SD)=6.5 (NR) y PL=NR WS=44% working PM=NR RMDQ, $\overline{X}$ (SD)=14.0 (5.4) ODI, $\overline{X}$ (SD)=35.0 (20.2)
						Responsiveness	48	Sex=67% female Age, $\overline{X}$ (SD)=52.0 (NR) y PD, $\overline{X}$ (SD)=6.0 (NR) y PL=NR WS=46% working PM=NR RMDQ, $\overline{X}$ (SD)=11.0 (6.1) ODI, $\overline{X}$ (SD)=29.0 (20.0)
Monticone et al,[44] 2012	Italy (Italian)	Rehabilitation unit of a hospital and 3 physical therapy services	Observational, longitudinal	Rehabilitation program (including exercises and cognitive-behavioral guidance)	NSLBP >6 wk (unspecified whether with or without leg pain)	Measurement error	102	Sex=NR Age, $\overline{X}$ (SD)=NR PD, $\overline{X}$ (SD)=NR PL=NR WS=NR PM=NR RMDQ, $\overline{X}$ (SD)=6.8 (4.6) ODI, $\overline{X}$ (SD)=27.7 (17.6)
						Construct validity Responsiveness	179	Sex=63% female Age, $\overline{X}$ (SD)=47.7 (12.3) y PD, $\overline{X}$ (SD)=6.0 (4.0) mo PL=NR WS=78% working PM=28% taking RMDQ, $\overline{X}$ (SD)=6.4 (4.3) ODI, $\overline{X}$ (SD)=26.8 (16.6)

(Continued)

Table 1.
Continued

Study	Country (Language)	Setting	Study Design	Health Intervention	Study Population	Measurement Properties	n	Patients' Characteristics
Mousavi et al,[45] 2006	Iran (Persian)	Physical therapy unit of a hospital	Observational, cross-sectional	NA	NSLBP (unspecified whether with or without leg pain)	Reliability	31	Sex=NR Age=NR PD=NR PL=NR WS=NR PM=NR RMDQ=NR ODI=NR
						Internal consistency Construct validity	100	Sex=55% female Age, $\bar{X}$ (SD)=40.1 (11.6) y PD, $\bar{X}$ (SD)=7.0 (8.8) y PL=NR WS=NR PM=NR RMDQ, $\bar{X}$ (SD)=9.7 (4.8) ODI, $\bar{X}$ (SD)=30.1 (12.4)

[a] NSLBP=nonspecific low back pain; PD=pain duration; PL=pain localization; WS=work status; PM=pain medication; RMDQ=Roland-Morris Disability Questionnaire, score at baseline; ODI=Oswestry Disability Index, score at baseline; GP=general practitioner; NA=not applicable to the design of the study; NR=not reported.
[b] These studies were included in the same article.
[c] These studies were included in the same article.

in the meta-analyses included patients with acute NSLBP, and its correlations were in line with the other studies in chronic NSLBP included in the meta-analyses; the only difference was a lower correlation with the SF-36 bodily pain subscale, but that was not substantially different between the RMDQ and ODI. Correlations in the meta-analyses were all as hypothesized (eAppendix 2) for the RMDQ (100%), whereas the ODI met 3 out of 5 of expected correlations (60%) (Figs. 2D and 2F).

In assessing the results of individual studies, the RMDQ met 62.5% of the a priori hypotheses, and the ODI met 75% of them (eTable). In performing best evidence synthesis, more weight was allocated to the results of the meta-analyses, as the meta-analyses were based on more precise correlation estimates. A moderate level of evidence with a consistent positive rating was given to the RMDQ, as all a priori hypotheses were met in the meta-analyses, whereas results for the ODI were considered as conflicting. These results indicate that the RMDQ has better construct validity than the ODI for measuring physical functioning in patients with NSLBP (Tab. 4).

Responsiveness

Seven studies[37–39,41,43,44] of fair methodological quality compared the responsiveness of the 2 instruments (Tab. 3). All but one study[43] assessed responsiveness using a construct approach, and all of the studies assessed responsiveness using a criterion approach, with a global perception of change scale (GPCS) as a gold standard and with an inconsistent number of point scales across studies (Tab. 3). The overall quality score of all studies[37–39,41,43,44] was influenced by different factors: unclear description of handling of missing items, vague or absent hypotheses regarding correlations or effect sizes, limited information on measurement properties of comparator instruments, and uncertainty regarding the GPCS as an adequate gold standard. Statistical pooling was not performed due to discrepancies in time points of assessment and differences in the GPCSs used in the different studies (Tab. 3). Like done for construct validity, disease-specific comparator instruments

Table 2.
Internal Consistency, Reliability, and Measurement Error of the RMDQ and ODI in Head-to-Head Comparison Studies Conducted in Patients With NSLBP[a]

Study	Study Population	Country (Language)	Internal Consistency				Reliability				Measurement Error			
			n	COSMIN	AD	Cronbach Alpha	n	COSMIN	Time Interval	ICC[b]	n	COSMIN	Time Interval	SEM[c] (% Scale Range)	SDC[d] (% Scale Range)
Davidson and Keating,[38] 2002	NSLBP	Australia (English)					47	Fair	6 wk	RMDQ=.53 ODI=.84	47	Fair	6 wk	RMDQ=3.7 (15%) ODI=6.0 (6%)	RMDQ=8.6 (36%) ODI=15.0 (15%)
Grotle et al,[40] 2003	Chronic NSLBP	Norway (Norwegian)	28	Poor	No	RMDQ=.94 ODI=.94	28	Poor	2 d	RMDQ=.89 ODI=.88	28	Poor	2 d	RMDQ=1.4 (6%)[e] ODI=4.0 (4%)[e]	RMDQ=4.0 (17%) ODI=11.0 (11%)
Maughan and Lewis,[43] 2010	Chronic NSLBP	United Kingdom (English)					25	Poor	5 wk	RMDQ=.90 ODI=.91[e]	25	Poor	5 wk	RMDQ=1.8 (7%) ODI=6.1 (6%)	RMDQ=4.9 (20%) ODI=16.7 (17%)
Monticone et al,[44] 2012	Subacute and chronic NSLBP	Italy (Italian)									102	Fair	8 wk	RMDQ=1.8 (7%)[e] ODI=4.9 (5%)[e]	RMDQ=4.9 (20%) ODI=13.7 (14%)
Mousavi et al,[45] 2006	NSLBP	Iran (Persian)	100	Poor	No	RMDQ=.83 ODI=.75	31	Fair	24 h	RMDQ=.86 ODI=.91					

[a] COSMIN=methodological quality assessment according to the COSMIN checklist, AD=assessment of dimensionality, ICC=intraclass correlation coefficient, SEM=standard error of measurement, SDC=smallest detectable change, NSLBP=nonspecific low back pain, RMDQ=Roland-Morris Disability Questionnaire, ODI=Oswestry Disability Index.
[b] Different ICC models were used (ie, Davidson and Keating[38]: ICC [2,1]; Grotle et al[40]: ICC [1,1]; Maughan and Lewis[43] and Mousavi et al[45]: not clear which model or formula was used).
[c] Not clear whether $SEM_{consistency}$ or $SEM_{agreement}$ was used.
[d] Different SDC formulas were used (ie, Davidson and Keating[38]: $\sqrt{}[2\times SEM]\times 1.64$; Grotle et al,[40] Maughan and Lewis,[43] and Monticone et al[44]: $SEM\times 2.77$).
[e] These values are not presented in the review, but we calculated them from the data available.

measuring function or disability were considered as measuring a similar construct, with all other instruments measuring other constructs as unrelated. In 3 studies,[41,44] correlations of change scores of both instruments with change scores in the SF-36-PF were lower than correlations with changes in pain measures, which was unexpected. In 2 of these studies,[41,44] the correlations with the SF-36-PF were below .5, and with some pain measurements above .50, both also were unexpected (Tab. 3). Forty percent of the correlations were in accordance with our hypotheses for the RMDQ, and 50% of the correlations were in accordance with our hypotheses for the ODI. In 6 studies,[37–39,41,44] both RMDQ and ODI displayed larger standardized response means (SRMs) for the group of "improved" patients when compared with the whole group or with those "not improved" (Tab. 3). In one study,[43] both questionnaires displayed areas under the curve (AUCs) below 0.70; in 2 studies,[39,44] only the RMDQ presented an AUC slightly below this threshold for a positive rating (eAppendix 2). The only study including solely patients with acute NSLBP[41] showed higher correlations, effect sizes, and AUCs than other studies, but the results were similar and conflicting for both questionnaires. Overall, due to a negative rating for correlations and a positive rating on SRMs (Tab. 3, eAppendix 2), the evidence was considered as conflicting for both instruments and consequently made inconclusive the comparison of responsiveness of the 2 instruments (Tab. 4).

Discussion

A systematic review was conducted to assess studies directly comparing the measurement properties of the original 24-item version of the RMDQ and version 2.1a of the ODI in patients with NSLBP. Nine articles, including 11 studies in the review, met the eligibility criteria (Fig. 1). There was moderate-quality evidence showing that the ODI has better test-retest reliability and less measurement error than the RMDQ (Tab. 4). On the other hand, there was moderate-quality evidence suggesting that the RMDQ has better construct validity than the ODI as a tool to assess physical functioning. Conflicting evidence was found

Table 3.
Responsiveness of the RMDQ and ODI in Studies Making a Head-to-Head Comparison in Patients With NSLBP[a]

Study	Population	Country (Language)	Health Intervention	Time Interval	n	COSMIN	Construct Approach		Criterion Approach		
							Correlations of Changes With Changes in Other Measures	SRMs[b]	Criterion	Correlations With Criterion	AUC
Coelho et al,[37] 2008	Chronic NSLBP	Brazil (Brazilian Portuguese)	Physical therapy (including manual therapy, exercises, physical conditioning, and electrotherapy)	6 wk	30	Fair		RMDQ total=0.44 RMDQ improved=0.89 RMDQ not improved=0.18 ODI total=0.37 ODI improved=0.78 ODI not improved=0.08	7-point GPCS		RMDQ=0.82 ODI=0.73
Davidson and Keating,[38] 2002	NSLBP	Australia (English)	Physical therapy	6 wk	106	Fair		RMDQ total=0.55 RMDQ improved=0.95[c] RMDQ not improved=0.16[c] ODI total=0.52 ODI improved=0.89[c] ODI not improved=0.11[c]	7-point GPCS		RMDQ=0.77 ODI=0.78
Frost et al,[39] 2008	Subacute and chronic NSLBP	United Kingdom (English)	Physical therapy plus one advice session or advice session alone	12 mo	201	Fair		RMDQ better=–0.70 RMDQ same=–0.07 RMDQ worse=1.25 ODI better=–0.86 ODI same=–0.01 ODI worse=0.61	3-point GPCS	RMDQ=0.38 ODI=0.47	RMDQ=0.69 ODI=0.75
Grotle et al,[41] 2004[d]	Acute NSLBP	Norway (Norwegian)	Information according to guidelines for acute LBP	4 wk	52	Fair	RMDQ vs SF-36-PF=.64 ODI vs SF-36-PF=.56 RMDQ vs NRS=.68 ODI vs NRS=.58 RMDQ vs DRI=.75 ODI vs DRI=.78	RMDQ total=0.9 RMDQ improved=1.5 RMDQ not improved=0.3 ODI total=0.9 ODI improved=1.4 ODI not improved=0.7	6-point GPCS	RMDQ=0.74 ODI=0.66	RMDQ=0.93 ODI=0.97
Grotle et al,[41] 2004[d]	Chronic NSLBP	Norway (Norwegian)	Treatment according to Indahl model for chronic LBP	3 mo	48	Fair	RMDQ vs SF-36-PF=.23 ODI vs SF-36-PF=.38 RMDQ vs NRS=.52 ODI vs NRS=.42 RMDQ vs DRI=.49 ODI vs DRI=.43	RMDQ total=0.2 RMDQ improved=1.1 RMDQ not improved=–0.1 ODI total=0.4 ODI improved=1.0 ODI not improved=0.2	6-point GPCS	RMDQ=0.61 ODI=0.49	RMDQ=0.83 ODI=0.75
Maughan and Lewis,[43] 2010	Chronic NSLBP	United Kingdom (English)	Physical therapy (including exercises and education)	5 wk	48	Fair			7-point GPCS		RMDQ=0.64 ODI=0.67

(Continued)

Table 3.
Continued

Study	Population	Country (Language)	Health Intervention	Time Interval	n	COSMIN	Construct Approach		Criterion Approach		
							Correlations of Changes With Changes in Other Measures	SRMs[b]	Criterion	Correlations With Criterion	AUC
Monticone et al,[44] 2012	Subacute and chronic NSLBP	Italy (Italian)	Rehabilitation program (including exercises and cognitive-behavioral guidance)	8 wk	179	Fair	RMDQ vs SF-36-PF=–.40 ODI vs SF-36-PF=–.40 RMDQ vs VAS=.47 ODI vs VAS=.54 RMDQ vs SF-36-BP=–0.52 ODI vs SF-36-BP=–0.48 RMDQ vs SF-36-RP=–0.37 ODI vs SF-36-RP=–0.33	RMDQ total=0.81 RMDQ improved=0.98[c] RMDQ not improved=–0.70[c] ODI total=0.80 ODI improved=1.13[c] ODI not improved=–0.60[c]	5-point GPCS	RMDQ=0.29 ODI=0.43	RMDQ=0.64 ODI=0.71

[a] COSMIN=methodological quality assessment according to the COSMIN checklist, AUC=area under the curve (receiving operating characteristic analysis), LBP=low back pain, NSLBP=nonspecific low back pain, RMDQ=Roland-Morris Disability Questionnaire, ODI=Oswestry Disability Index, DRI=Disability Rating Index, SF-36-PF=physical functioning subscale of the 36-Item Short-Form Health Survey, NRS=numeric rating scale, VAS=visual analog scale, SF-RP-36=role–physical subscale of the 36-Item Short-Form Health Survey, SF-36-BP=bodily pain subscale of the 36-Item Short-Form Health Survey, GPCS=global perception of change scale, SRM=standardized response mean.
[b] Standardized response means were calculated by dividing the mean change by its standard deviation.
[c] These SRMs were calculated using the data presented in the review.
[d] These studies were included in the same article.

for responsiveness of both instruments, and their internal consistency is unknown due to only studies of methodological quality or no studies on that measurement property (Tab. 4). In this review, no clearly different findings in measurement properties could be shown for patients with a different NSLBP duration. Overall, based on the 5 measurement properties assessed in the studies included in this review, there are no strong arguments to prefer one instrument over the other to measure physical functioning in patients with NSLBP. Nevertheless, this systematic review provides some valuable information that should be put in the research agenda by the scientific community. First, head-to-head comparison studies of adequate methodological quality on the 5 measurement properties included in this review are necessary. Second, and more importantly, some key measurement properties of these 2 instruments (ie, content, structural, and cross-cultural validity) should be compared in patients with NSLBP.

The ODI 2.1a was found to have better test-retest reliability, mainly because an ICC below .70 was found for the 24-item RMDQ in one study[38] of fair methodological quality (Tab. 2). A recent systematic review[23] retrieved 28 studies assessing test-retest reliability of all RMDQ versions, finding only 2 studies displaying an ICC below .70.[38,56] These results might suggest that the results of the study by Davidson and Keating[38] could be considered as fortuitous or strictly related to the long time interval used for reassessment; the same aforementioned review also found that heterogeneity in reliability results across studies can be explained by the different test-retest time frames adopted.[23] However, the same review also substantiated the results for test-retest reliability found in this review, as the pooled ICC for the ODI was higher than that of the RMDQ.[23]

The studies included in this review showed that the ODI has a smaller measurement error than the RMDQ, also explaining the results in favor of the ODI for test-retest reliability. These findings are in line with those of the review of Geere et al,[23] who found smaller mean

SDCs for the ODI when all versions of the 2 questionnaires were considered. However, in that review,[23] the difference of SDC of the 2 questionnaires was not present when only time intervals shorter than 14 days were analyzed. This difference could not be assessed in our review because the 2 studies of fair methodological quality[38,44] adopted time intervals of 6 and 8 weeks, respectively. Another way to assess the measurement error of an instrument is to compare it with its MIC and evaluate whether the instrument is able to discriminate measurement error from the MIC.[36] Nevertheless, a limitation of this approach is that it might be difficult to use absolute MIC values for an instrument, considering that they can be context- and population-specific[57] and dependent on baseline values of the assessed questionnaires.[58] Two of the studies included in this review[43,44] also estimated the MIC of the 2 instruments and showed them to be smaller than SDCs for both questionnaires. This finding would indicate that neither of the 2 instruments is able to discriminate between SDC and MIC or to detect "real" changes in the construct measured. Hence, although the ODI has a smaller measurement error than the RMDQ, it cannot be asserted that it also has a greater ability in discriminating SDC and MIC.

The ODI consistently displayed higher correlations with other instruments measuring the same or unrelated constructs (eg, pain intensity, general health, mental health, social functioning) (Fig. 2, eTable). On the one hand, these results could suggest that the construct measured by the 2 questionnaires is not precisely the same and that the construct measured by the ODI might be broader than that of the RMDQ; they also might indicate that the RMDQ measures a narrower construct and that it might provide a more focused assessment of physical functioning. On the other hand, it also is possible that the stronger correlations of the ODI could be partly explained by its smaller measurement error documented in this review. It should be noted that we made the a priori decision to consider pain intensity as an unrelated construct because the purpose of our study was to assess RMDQ and ODI as measures of physical functioning, defined as "the ability to carry out daily physical activities."[8(p1133)] This subjective decision has strongly influenced the specific conclusion that the RMDQ has better construct validity and the general conclusion that there are no strong arguments to prefer 1 of the 2 instruments. This subjective decision could be criticized, as pain intensity also could be considered as a construct related to RMDQ and ODI, as they are LBP-specific instruments. These 2 instruments were developed to measure disability,[14,15] which, taking into account frequently used models and definitions,[59,60] is a domain that cannot be considered equivalent to physical functioning as defined for this study. Moreover, previous analyses of the content of the 2 instruments have shown that they do not measure only daily physical activities.[9,61,62] Overall, the results of this review on construct validity should be further explored by future studies of good or excellent methodological quality formulating multiple and specific a priori hypotheses regarding expected correlations with other instruments and by studies comparing the content validity of the 2 instruments as measures of physical functioning or of a larger construct.

Responsiveness was assessed by the majority of the studies included in this review, but conflicting evidence was found for both the RMDQ and ODI

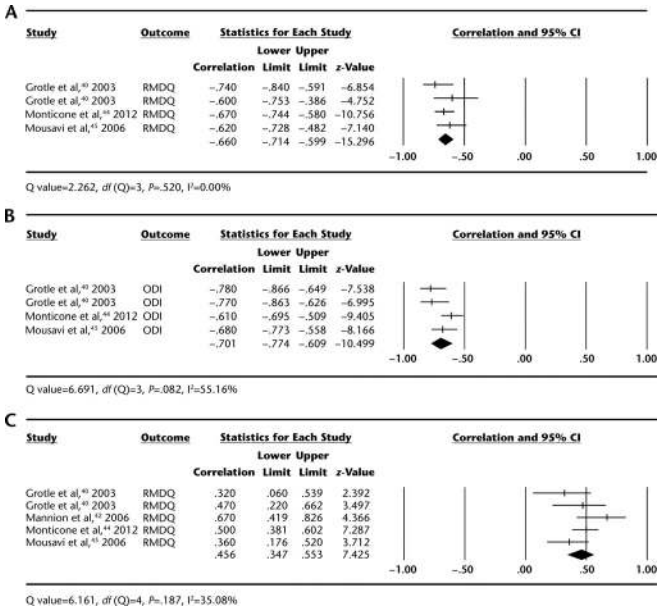

Figure 2.
Pooled correlations with 95% confidence intervals (95% CIs) of Roland-Morris Disability Questionnaire (RMDQ) and Oswestry Disability Index (ODI) with other instruments measuring related or unrelated constructs in patients with nonspecific low back pain: (A) correlation between RMDQ and physical functioning subscale of the 36-Item Short-Form Health Survey (SF-36-PF) in 384 patients, (B) correlation between ODI and SF-36-PF in 384 patients, (C) correlation between RMDQ and pain intensity measures (pain intensity was measured with a 100-mm visual analogue scale by Grotle et al,[40] Mannion et al,[42] and Mousavi et al[45] and with a 0–10 numeric rating scale by Monticone et al[44]) in 416 patients, (D) correlation between ODI and pain intensity measures (pain intensity was measured with a 100-mm visual analogue scale by Grotle et al,[40] Mannion et al,[42] and Mousavi et al[45] and with a 0–10 numeric rating scale by Monticone et al[44]) in 416 patients, (E) correlation between RMDQ and bodily pain subscale of the 36-Item Short-Form Health Survey (SF-36-BP) in 384 patients, (F) correlation between ODI and SF-36-BP in 384 patients, (G) correlation between RMDQ and general health subscale of the 36-Item Short-Form Health Survey (SF-36-GH) in 205 patients, (H) correlation between ODI and SF-36-GH in 205 patients, (I) correlation between RMDQ and mental health subscale of the 36-Item Short-Form Health Survey (SF-36-MH) in 205 patients, and (J) correlation between ODI and SF-36-MH in 205 patients.

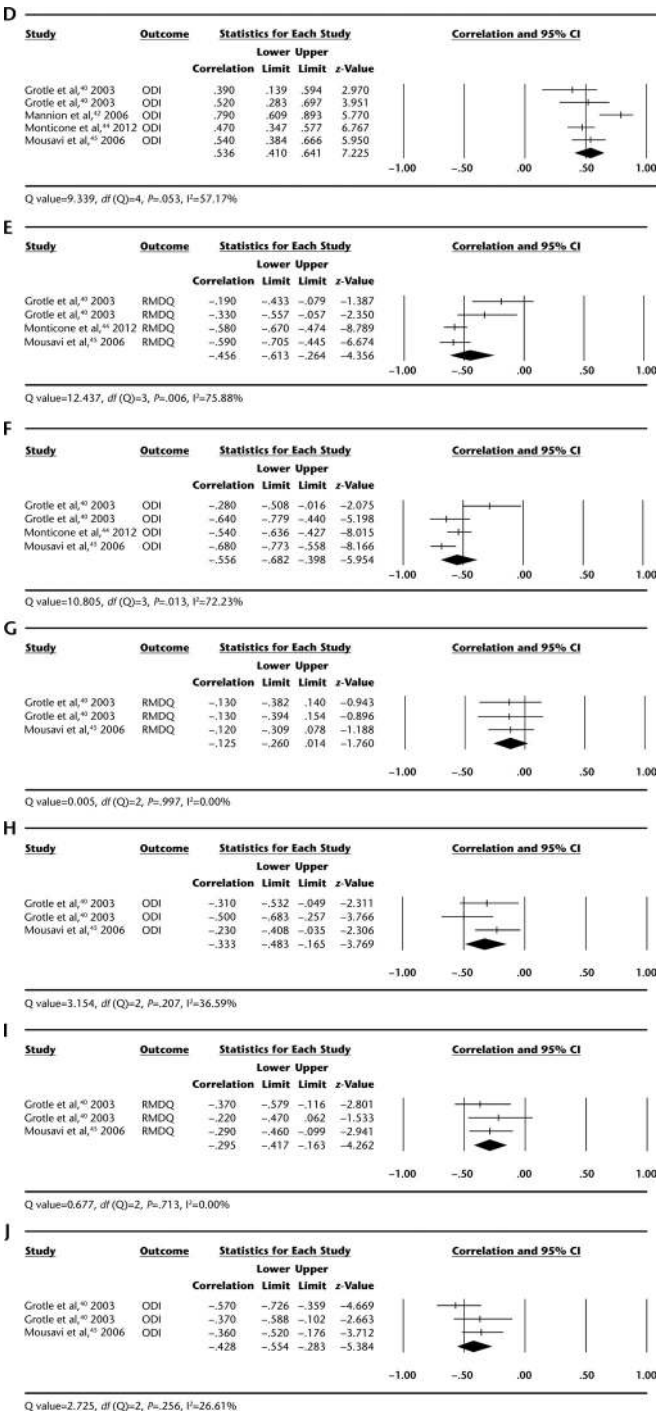

Figure 2.
Continued

(Tab. 4). All of the studies lacked the formulation of multiple a priori and specific hypotheses regarding expected correlations with changes in other instruments or effect sizes; this gap should be filled by future longitudinal studies assessing this measurement property.[20] Another methodological aspect that should be improved in future studies is the formulation of GPCSs used as gold standards to assess responsiveness following a criterion approach.[20] It was recently shown that construct-specific anchors have higher validity than global anchors,[63] as those used in the studies included in this review used generic GPCSs. Considering that both the RMDQ and ODI are widely used as outcome measurement instruments,[10] it is fundamental that they display good responsiveness, and studies of good or excellent quality are needed to better assess this measurement property. The rating of conflicting evidence in this review also was driven by the fact that correlations between the change on the RMDQ or ODI and the change on the SF-36-PF were found to be lower than correlations obtained with instruments not measuring the same construct (Tab. 3). It could easily be asserted that these unexpected results were due to the poor responsiveness of the SF-36-PF, but more studies have shown that, besides displaying lower results than other measurement tools, the responsiveness of the SF-36-PF is above minimum criteria for both AUCs and effect sizes.[64–67]

The comparison of internal consistency of the 2 instruments was inconclusive, considering that unidimensionality of the instruments was not assessed. Despite the fact that these are the 2 most widely used outcome measurement instruments,[10] no studies comparing the structural validity of these questionnaires were found. In recent years, some studies assessed the dimensionality of one or the other instrument by means of factor analysis or Rasch analysis. Results of these studies are contradictory regarding the dimensionality of both questionnaires, with some studies revealing them to be unidimensional[68–73] and others not.[40,74–76] The results of these studies highlight that it is not clear whether both instruments are unidimensional and that, possibly, their internal structure might vary across different languages and populations. For this reason, it is crucial that future studies on the RMDQ and ODI compare their structural validity in the same population and that they do so before assessing their internal consistency. It also is suggested to further evaluate cross-cultural validity of both questionnaires, as this evaluation will give insight into possible differences in factor

Table 4.
Best Evidence Synthesis of Measurement Properties of the RMDQ and ODI in Head-to-Head Comparison Studies Conducted in Patients With Nonspecific Low Back Pain[a]

Measurement Properties	RMDQ Level of Evidence (Rating)	ODI Level of Evidence (Rating)	Is One Instrument Better Than the Other?
Internal consistency	?	?	?
Reliability	Conflicting (+/–)	Moderate (+)	Yes, ODI
Measurement error	Moderate (–)	Moderate (+)	Yes, ODI
Face validity	?	?	?
Content validity	?	?	?
Structural validity	?	?	?
Construct validity	Moderate (+)	Conflicting (+/–)	Yes, RMDQ
Cross-cultural validity	?	?	?
Criterion validity	?	?	?
Responsiveness	Conflicting (+/–)	Conflicting (+/–)	No

[a] RMDQ=Roland-Morris Disability Questionnaire, ODI=Oswestry Disability Index, ?=unknown due to only studies of poor methodological quality or no studies on that measurement property, +/–=conflicting findings, +=consistent positive findings, –=consistent negative findings.

structure or differential item functioning across translated versions.

Studies performing a head-to-head comparison of the content validity of recommended versions of the RMDQ and ODI are needed to evaluate whether the content of 1 of the 2 instruments better represents the most relevant aspects of physical functioning in patients with NSLBP. Content validity refers to the extent to which, in a given measurement application, the most relevant and comprehensive aspects of a construct are adequately reflected in the content of a measurement instrument.[19,77] To date, the RMDQ and ODI have been considered and recommended for measuring the same health construct.[11–13] However, the results on construct validity of our review clearly suggest that there are discrepancies in their correlations with other instruments and that, possibly, they do not measure exactly the same construct. Two studies comparing the content of RMDQ and ODI showed that some body functions or activity limitations were related to the items of one questionnaire but not the other, and vice versa.[61,62] Considering the emerging importance of this measurement property in the selection of instruments and in the assessment of their validity,[77–80] it is essential to investigate content validity further for these 2 questionnaires. In general, when making a choice between 2 instruments, content validity should be the first property to be explored to evaluate whether one instrument is a better reflection of the construct measured in the specific target population.

Here, we provide some suggestions for future studies assessing and comparing content, structural, and cross-cultural validity of the RMDQ and ODI in patients with NSLBP. Qualitative research plays a key role in the assessment of content validity of existing instruments.[81] Focus groups or cognitive interviews[81] could be conducted in patients with NSLBP to assess which of the 2 instruments cover the most important aspects of physical functioning and whether there are additional relevant aspects that are not covered. Previous studies have assessed the content of these instruments by linking it to the *International Classification of Functioning* (ICF) categories.[61,62] However, to date, no studies have attempted to link the content of the RMDQ and ODI to the ICF core set for LBP.[82] Focusing only on the ICF categories included in the core set would allow us to better understand whether the content of these instruments reflects and covers several aspects important to patients with LBP. A recent study that followed this procedure with the ICF core set for rheumatoid arthritis could be used as a valid example for such a study.[83] The qualitative assessment of content validity of both instruments should be combined with the quantitative assessment of their structural validity (ie, dimensionality).[77] Evidence on the unidimensionality of the RMDQ and ODI should be provided, as their total score is routinely used to assess the effectiveness of health interventions in patients with NSLBP.[84]

Statistical techniques such as confirmatory factor analysis[85] or item response theory (IRT) models[86,87] allow us to assess the unidimensionality of a patient-reported instrument. Item response theory models provide some advantages over factor analysis,[88,89] as they also permit us to estimate item parameters along a continuum representing different levels of ability on the measured construct and to estimate the measurement precision of an instrument along the same continuum.[86,87] Therefore, IRT should be preferred to compare the performance of the RMDQ and ODI in the same group of patients with NSLBP, but authors of future IRT studies should be aware that, when testing both instruments in the same patients, a large sample size is required. For example, if using an IRT 2-parameter logistic model (eg, graded response model), at least 1,000 participants are needed to have accurate parameters' estimation.[87,89]

To date, to our knowledge, there are no studies examining whether the factor structure of the RMDQ and ODI is consistent across different countries and languages, and, for this reason, it is of high priority to assess *cross-cultural validity*, defined as "the degree to which the performance of the items on a translated or culturally adapted patient-reported instrument are an adequate reflection of the performance of the items of the original version of the instrument."[19(p743)] It should be highlighted that cross-cultural validity refers not only to the factor structure of a questionnaire but also to other aspects of validity, such as face and content validity. Therefore, it would be important that cross-cultural adaptation processes include an assessment of face and content validity in patients with NSLBP, as these properties can vary for the same instrument in different languages, cultures, and settings. Having empirical evidence supporting the cross-

cultural validity of the RMDQ and ODI would allow reviewers to combine studies with more confidence in future systematic reviews. A study on cross-cultural validity of these widely used instruments would require a collaborative effort of the scientific community to join forces and design parallel studies in different countries. Such an effort could be facilitated or embedded within already active collaborations such as the international and multidisciplinary steering committee working on the development of a core outcome set for clinical trials in NSLBP.[8,27]

It was out of the scope of this review to compare the RMDQ and ODI as measures of disability, but our results give a clear indication on this matter in patients with NSLBP. If researchers or clinicians want to measure a functional domain broader than solely physical functioning, the ODI should be preferred over the RMDQ because: (1) it displays better test-retest reliability and measurement error and (2) the higher cross-sectional correlations with all other instruments would indicate better construct validity. As previously reported, consistent higher correlations with other instruments can be explained by the fact that the instrument measures a broader construct or has smaller measurement error, which would support the preference for the ODI in both cases.

A previous review[13] recommended use of the ODI in patients with persistent and severe disability and use of the RMDQ in patients with lower levels of disability. These recommendations were based on a previous study[48] showing differences related to floor and ceiling effects, with the greater proportion of patients scoring higher on the RMDQ or lower on the ODI. All studies included in this review reported very similar score levels on the RMDQ and ODI (Tab. 1), making it not feasible to empirically assess these previous recommendations. However, we attempted to assess whether a difference in some measurement properties could be related to the pain duration (ie, acute versus subacute or chronic). Only 2 studies[40,41] in patients with acute NSLBP were included, and they did not show a substantial different trend in results between the 2 questionnaires. Hence, due to the small amount of head-to-head comparisons in acute NSLBP, we are not able make suggestions regarding one instrument being better (or not) than the other in patients with a different pain duration.

Three considerations can be made on some methodological aspects of this review. First, we chose to focus on RMDQ and ODI versions recommended by their developers[13] because they showed superior measurement properties compared with shorter or modified versions in patients with NSLBP.[65,68,69,90] Consequently, the results of this review cannot be generalized to all existing versions of these questionnaires. Second, the results of this review are not generalizable to specific LBP populations (eg, patients with spinal stenosis), as we focused on and included only studies conducted in patients with NSLBP; this decision was taken to be consistent with the scope of the core outcome set that has been developed for clinical trials in patients with NSLBP.[8,27] Third, a potential limitation of this study is that we combined RMDQ and ODI data from different countries and languages without knowing whether the items of these questionnaires have the same performance in different cultures. Hence, the evaluation of validity of these questionnaires across different languages and countries should have very high priority. This is an important issue not only for using these patient-reported outcomes in clinical practice but also in systematic reviews in which data from different cultures and languages are routinely combined.

To sum up, this systematic review identified 11 studies of fair or poor methodological quality, performing head-to-head comparisons of 5 measurement properties of the 24-item RMDQ and ODI 2.1a. The ODI showed better reliability and measurement error, whereas the RMDQ showed better construct validity as a measure of physical functioning. In light of these findings, there are no strong reasons to prefer one instrument over the other to measure physical functioning in patients with NSLBP. To further compare the measurement properties of these 2 instruments, studies of higher methodological quality are needed, and priority should be given to studies on the content, structural, and cross-cultural validity of these questionnaires.

Mr Chiarotto, Ms Maxwell, Dr Terwee, Professor Wells, Professor Tugwell, and Professor Ostelo provided concept/idea/research design. Mr Chiarotto, Dr Terwee, and Professor Ostelo provided project management. Mr Chiarotto and Ms Maxwell provided data collection. Mr Chiarotto, Ms Maxwell, Dr Terwee, and Professor Ostelo provided data analysis. Professor Ostelo provided fund procurement. All authors provided writing and consultation (including review and approval of manuscript before submission).

The authors acknowledge the Wetenshcappelijk College Fysiotherapie (WCF) of the Royal Dutch Society for Physical Therapy (KNGF) for providing funding for this study.

DOI: 10.2522/ptj.20150420

References

1 Vos T, Flaxman AD, Naghavi M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet.* 2013;380:2163–2196.

2 Hoy D, Brooks P, Blyth F, Buchbinder R. The epidemiology of low back pain. *Best Pract Res Clin Rheumatol.* 2010;24:769–781.

3 Dagenais S, Caro J, Haldeman S. A systematic review of low back pain cost of illness studies in the United States and internationally. *Spine J.* 2008;8:8–20.

4 Lambeek LC, van Tulder MW, Swinkels IC, et al. The trend in total cost of back pain in The Netherlands in the period 2002 to 2007. *Spine (Phila Pa 1976).* 2011;36:1050–1058.

5 Balagué F, Mannion AF, Pellisé F, Cedraschi C. Non-specific low back pain. *Lancet.* 2012;379:482–491.

6 Cohen SP, Argoff CE, Carragee EJ. Management of low back pain. *BMJ.* 2008;337:a2718.

7 Koes B, van Tulder M, Thomas S. Diagnosis and treatment of low back pain. *BMJ.* 2006;332:1430–1434.

8 Chiarotto A, Deyo RA, Terwee CB, et al. Core outcome domains for clinical trials in non-specific low back pain. *Eur Spine J.* 2015;24:1127–1142.

9 Grotle M, Brox JI, Vøllestad NK. Functional status and disability questionnaires: what do they assess? A systematic review of back-specific outcome questionnaires. *Spine (Phila Pa 1976).* 2005;30:130–140.

10 Chapman JR, Norvell DC, Hermsmeyer JT, et al. Evaluating common outcomes for measuring treatment success for chronic low back pain. *Spine (Phila Pa 1976).* 2011;36:S54–S68.

11 Bombardier C. Outcome assessments in the evaluation of treatment of spinal disorders: summary and general recommendations. *Spine (Phila Pa 1976).* 2000;25:3100–3103.

12 Deyo RA, Battie M, Beurskens A, et al. Outcome measures for low back pain research: a proposal for standardized use. *Spine (Phila Pa 1976).* 1998;23:2003–2013.

13 Roland M, Fairbank J. The Roland–Morris Disability Questionnaire and the Oswestry Disability Questionnaire. *Spine (Phila Pa 1876).* 2000;25:3115–3124.

14 Roland M, Morris R. A study of the natural history of back pain, part I: development of a reliable and sensitive measure of disability in low-back pain. *Spine (Phila Pa 1976).* 1983;8:141–144.

15 Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. *Physiotherapy.* 1980;66:271–273.

16 Fairbank JC, Pynsent PB. The Oswestry Disability Index. *Spine (Phila Pa 1976).* 2000;25:2940–2953.

17 Baker D, Pynsent P, Fairbank JC. The Oswestry Disability Index revisited: its reliability, repeatability and validity, and a comparison with the St Thomas's Disability Index. In Roland MO, Jenner JR, eds. *Back Pain: New Approaches to Rehabilitation and Education.* Manchester, United Kingdom: Manchester University Press; 1989:174–186.

18 Fairbank JC. Why are there different versions of the Oswestry Disability Index? A review. *J Neurosurg Spine.* 2014;20:83–86.

19 Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol.* 2010;63:737–745.

20 de Vet HC, Terwee CB, Mokkink LB, Knol DL. *Measurement in Medicine: A Practical Guide*. Cambridge, United Kingdom: Cambridge University Press; 2011.

21 Cleland J, Gillani R, Bienen EJ, Sadosky A. Assessing dimensionality and responsiveness of outcomes measures for patients with low back pain. *Pain Pract.* 2011;11:57–69.

22 Davies CC, Nitz AJ. Psychometric properties of the Roland-Morris Disability Questionnaire compared to the Oswestry Disability Index: a systematic review. *Phys Ther Rev.* 2009;14:399–408.

23 Geere JH, Geere J-AL, Hunter PR. Meta-analysis identifies Back Pain Questionnaire reliability influenced more by instrument than study design or population. *J Clin Epidemiol.* 2013;66:261–267.

24 Smeets R, Köke A, Lin CW, et al. Measures of function in low back pain/disorders: Low Back Pain Rating Scale (LBPRS), Oswestry Disability Index (ODI), Progressive Isoinertial Lifting Evaluation (PILE), Quebec Back Pain Disability Scale (QBPDS), and Roland-Morris Disability Questionnaire (RDQ). *Arthritis Care Res.* 2011;63(suppl 11):S158–S173.

25 Fritz JM, Irrgang JJ. A comparison of a modified Oswestry Low Back Pain Disability Questionnaire and the Quebec Back Pain Disability Scale. *Phys Ther.* 2001;81:776–788.

26 Newman AN, Stratford PW, Letts L, Spadoni G. A Systematic review of head-to-head comparison studies of the Roland-Morris and Oswestry measures' abilities to assess change. *Physiother Can.* 2013;65:160–166.

27 Chiarotto A, Terwee CB, Deyo RA, et al. A core outcome set for clinical trials on nonspecific low back pain: study protocol for the development of a core domain set. *Trials.* 2014;15:511.

28 Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Ann Intern Med.* 2009;151:W-65–W-94.

29 Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med.* 2009;151:264–269.

30 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.* 2009;18:1115–1123.

31 Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res.* 2010;19:539–549.

32 Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res.* 2012;21:651–657.

33 Borenstein M, Hedges LV, Higgins JP, Rothstein HR. *Introduction to Meta-analysis*. New York, NY: John Wiley & Sons Inc; 2011.

34 DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials.* 1986;7:177–188.

35 Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* 2003;327:557–560.

36 Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol.* 2007;60:34–42.

37 Coelho RA, Siqueira FB, Ferreira PH, Ferreira ML. Responsiveness of the Brazilian-Portuguese version of the Oswestry Disability Index in subjects with low back pain. *Eur Spine J.* 2008;17:1101–1106.

38 Davidson M, Keating JL. A comparison of five low back disability questionnaires: reliability and responsiveness. *Phys Ther.* 2002;82:8–24.

39 Frost H, Lamb SE, Stewart-Brown S. Responsiveness of a patient-specific outcome measure compared with the Oswestry Disability Index v2. 1 and Roland and Morris Disability Questionnaire for patients with subacute and chronic low back pain. *Spine (Phila Pa 1976).* 2008;33:2450–2457.

40 Grotle M, Brox JI, Vøllestad NK. Cross-cultural adaptation of the Norwegian versions of the Roland-Morris Disability Questionnaire and the Oswestry Disability Index. *J Rehabil Med.* 2003;35:241–247.

41 Grotle M, Brox JI, Vøllestad NK. Concurrent comparison of responsiveness in pain and functional status measurements used for patients with low back pain. *Spine (Phila Pa 1976).* 2004;29:E492–E501.

42 Mannion A, Junge A, Fairbank J, et al. Development of a German version of the Oswestry Disability Index, part 1: cross-cultural adaptation, reliability, and validity. *Eur Spine J.* 2006;15:55–65.

43 Maughan EF, Lewis JS. Outcome measures in chronic low back pain. *Eur Spine J.* 2010;19:1484–1494.

44 Monticone M, Baiardi P, Vanti C, et al. Responsiveness of the Oswestry Disability Index and the Roland-Morris Disability Questionnaire in Italian subjects with subacute and chronic low back pain. *Eur Spine J.* 2012;21:122–129.

45 Mousavi SJ, Parnianpour M, Mehdian H, et al. The Oswestry Disability Index, the Roland-Morris Disability Questionnaire, and the Quebec Back Pain Disability Scale: translation and validation studies of the Iranian versions. *Spine (Phila Pa 1976).* 2006;31:E454–E459.

46 Beurskens A, de Vet HC, Köke A. Responsiveness of functional status in low back pain: a comparison of different instruments. *Pain.* 1996;65:71–76.

47 Boscainos PJ, Sapkas G, Stilianessi E, et al. Greek versions of the Oswestry and Roland-Morris Disability Questionnaires. *Clinical Orthop Rel Res.* 2003;411:40–53.

48 Kopec JA, Esdaile JM, Abrahamowicz M, et al. The Quebec Back Pain Disability Scale: Measurement Properties. *Spine (Phila Pa 1976).* 1995;20:341–352.

49 Leclaire R, Blier F, Fortin L, Proulx R. A cross-sectional study comparing the Oswestry and Roland-Morris Functional Disability scales in two populations of patients with low back pain of different levels of severity. *Spine (Phila Pa 1976).* 1997;22:68–71.

50 Reneman MF, Jorritsma W, Schellekens JM, Göeken LN. Concurrent validity of questionnaire and performance-based disability measurements in patients with chronic nonspecific low back pain. *J Occup Rehabil.* 2002;12:119–129.

51 Stratford PW, Binkley J, Solomon P, et al. Assessing change over time in patients with low back pain. *Phys Ther.* 1994;74:528–533.

52 Hsieh C, Phillips RB, Adams A, Pope M. Functional outcomes of low back pain: comparison of four treatment groups in a randomized controlled trial. *J Manipulative Physiol Ther.* 1992;15:4–9.

53 Lauridsen HH, Hartvigsen J, Manniche C, et al. Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients. *BMC Musculoskelet Disord.* 2006;7:82.

54 Mannion A, Junge A, Grob D, et al. Development of a German version of the Oswestry Disability Index, part 2: sensitivity to change after spinal surgery. *Eur Spine J.* 2006;15:66–73.

55 Moon J, Kim YC, Park SY, et al. Psychometric characteristics of the Korean version of the Roland-Morris Disability Questionnaire. *J Korean Med Sci.* 2011;26:1364–1370.

56 Dunn KM, Jordan K, Croft PR. Does questionnaire structure influence response in postal surveys? *J Clin Epidemiol.* 2003;56:10–16.

57 Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol.* 2008;61:102–109.

58 de Vet HC, Foumani M, Scholten MA, et al. Minimally important change values of a measurement instrument depend more on baseline values than on the type of intervention. *J Clin Epidemiol.* 2015;68:518–524.

59 Jette AM. Physical disablement concepts for physical therapy research and practice. *Phys Ther.* 1994;74:380–386.

60 *Towards a Common Language for Functioning, Disability and Health: ICF*. Geneva, Switzerland: World Health Organization; 2002.

61 Sigl T, Cieza A, Brockow T, et al. Content comparison of low back pain-specific measures based on the *International Classification of Functioning, Disability and Health* (ICF). *Clin J Pain.* 2006;22:147–153.

62 Wang P, Zhang J, Liao W, et al. Content comparison of questionnaires and scales used in low back pain based on the *International Classification of Functioning, Disability and Health*: a systematic review. *Disabil Rehabil.* 2012;34:1167–1177.

63 Ward MM, Guthrie LC, Alba M. Domain-specific transition questions demonstrated higher validity than global transition questions as anchors for clinically important improvement. *J Clin Epidemiol.* 2015;68:655–661.

64 Bronfort G, Bouter LM. Responsiveness of general health status in chronic low back pain: a comparison of the COOP charts and the SF-36. *Pain.* 1999;83:201–209.

65 Ostelo RW, de Vet HC, Knol DL, van den Brandt PA. 24-item Roland-Morris Disability Questionnaire was preferred out of six functional status questionnaires for post-lumbar disc surgery. *J Clin Epidemiol.* 2004;57:268–276.

66 Taylor SJ, Taylor AE, Foy MA, Fogg AJ. Responsiveness of common outcome measures for patients with low back pain. *Spine (Phila Pa 1976).* 1999;24:1805–1812.

67 Walsh TL, Hanscom B, Lurie JD, Weinstein JN. Is a condition-specific instrument for patients with low back pain/leg symptoms really necessary? The responsiveness of the Oswestry Disability Index, MODEMS, and the SF-36. *Spine (Phila Pa 1976).* 2003;28:607–615.

68 Davidson M. Rasch analysis of three versions of the Oswestry Disability Questionnaire. *Man Ther.* 2008;13:222–231.

69 Davidson M. Rasch analysis of 24-, 18- and 11-item versions of the Roland-Morris Disability Questionnaire. *Qual Life Res.* 2009;18:473–481.

70 Garratt AM, United Kingdom Back Pain Exercise and Manipulation Trial. Rasch analysis of the Roland Disability Questionnaire. *Spine (Phila Pa 1976).* 2003;28:79-84.

71 Mâaroufi H, Benbouazza K, Faïk A, et al. Translation, adaptation, and validation of the Moroccan version of the Roland-Morris Disability Questionnaire. *Spine (Phila Pa 1976).* 2007;32:1461–1465.

72 Valasek T, Varga PP, Szövérfi Z, et al. Reliability and validity study on the Hungarian versions of the Oswestry Disability Index and the Quebec Back Pain Disability Scale. *Eur Spine J.* 2013;22:1010–1018.

73 van Hooff ML, Spruit M, Fairbank JC, et al. The Oswestry Disability Index (version 2.1a): validation of a Dutch language version. *Spine (Phila Pa 1976).* 2015;40:E83–E90.

74 Algarni A, Ghorbel S, Jones J, Guermazi M. Validation of an Arabic version of the Oswestry Index in Saudi Arabia. *Ann Phys Rehabil Med.* 2014;57:653–663.

75 Pekkanen L, Kautiainen H, Ylinen J, et al. Reliability and validity study of the Finnish version 2.0 of the Oswestry Disability Index. *Spine (Phila Pa 1876).* 2011;36:332–338.

76 Magnussen LH, Lygren H, Strand LI, et al. Reconsidering the Roland-Morris Disability Questionnaire: time for a multidimensional framework? *Spine (Phila Pa 1976).* 2015;40:257–263.

77 Magasi S, Ryan G, Revicki D, et al. Content validity of patient-reported outcome measures: perspectives from a PROMIS meeting. *Qual Life Res.* 2012;21:739–746.

78 Ailliet L, Knol DL, Rubinstein SM, et al. Definition of the construct to be measured is a prerequisite for the assessment of validity: the Neck Disability Index as an example. *J Clin Epidemiol.* 2013;66:775–782. e772.

79 Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report, part 2: assessing respondent understanding. *Value Health.* 2011;14:978–988.

80 Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report, part 1—eliciting concepts for a new PRO instrument. *Value Health.* 2011;14:967–977.

81 Brod M, Tesler LE, Christensen TL. Qualitative research and content validity: developing best practices based on science and experience. *Qual Life Res.* 2009;18:1263–1278.

82 Cieza A, Stucki G, Weigl M, et al. ICF Core Sets for low back pain. *J Rehabil Med.* 2004;36:69–74.

83 Oude Voshaar MA, Ten Klooster PM, Glas CA, et al. Validity and measurement precision of the PROMIS physical function item bank and a content validity–driven 20-item short form in rheumatoid arthritis compared with traditional measures. *Rheumatology (Oxford).* 2015;54:2221–2229.

84 Kamper SJ, Apeldoorn A, Chiarotto A, et al. Multidisciplinary biopsychosocial rehabilitation for chronic low back pain: Cochrane systematic review and meta-analysis. *BMJ.* 2015;350:h444.

85 Thompson B. *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications*. Washington, DC: American Psychological Association; 2004.

86 DeMars C. *Item Response Theory*. Oxford, United KIngdom: Oxford University Press; 2010.

87 Embretson SE, Reise SP. *Item Response Theory*. Hove, United Kingdom: Psychology Press; 2013.

88 Petrillo J, Cano SJ, McLeod LD, Coon CD. Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples. *Value Health.* 2015;18:25–34.

89 Reeve BB, Fayers P. Applying item response theory modeling for evaluating questionnaire item and scale properties. In: Fayers P, Hays RD, eds. *Assessing Quality of Life in Clinical Trials: Methods of Practice*. 2nd ed. New York, NY: Oxford University Press; 2005:53–73.

90 Grotle M, Wilkens P, Garratt AM, et al. Which Roland-Morris Disability Questionnaire? Rasch analysis of four different versions tested in a Norwegian population. *J Rehabil Med.* 2013;45:670–677.