

[LITERATURE REVIEW]

JOY C. MACDERMID, PhD¹ • DAVID M. WALTON, MSc² • SARAH AVERY, MScPT³ • ALANNA BLANCHARD, MScPT³
EVELYN ETRUW, MScPT³ • CHERYL MCALPINE, MScPT³ • CHARLIE H. GOLDSMITH, PhD⁴

Measurement Properties of the Neck Disability Index: A Systematic Review

It is estimated that a third of all adults will experience neck pain during the course of 1 year,¹² and 70% is the approximate lifetime prevalence.^{6,28} About 19% of the population may suffer from chronic neck pain at any given time,⁶ creating a substantial societal burden. Monitoring outcomes is a key component of monitoring the effects of evidence-based care, in justifying services, and in program evaluation.

However, the episodic, fluctuating nature of neck pain and the lack of clear and consistent physiological findings complicate this process. A fundamental component of monitoring outcomes is having reliable and valid tools with known measurement properties.²⁷ In the case of neck disorders, self-reported pain and disability are usually the primary focus.

Isolated studies on outcomes measures provide a context and method-specific view of the measure's ability to provide useful clinical information. It is only by synthesizing information from multiple studies that we can understand how a measurement tool performs across different contexts and applications. The synthesis of larger pools of data is a mechanism to provide more stable estimates of measurement errors and benchmarks for change/outcomes. In fact, 2 systematic reviews have been conducted that address self-report measures for neck pain. Both compared different outcome measures using a semistructured process and concluded that the Neck Disability Index (NDI) was the most commonly used self-report instrument for evaluating status in neck pain clinical research.^{34,37} Both reviews alluded to the psychometric and clinical properties of the tools, but neither attempted to deal with specific properties or to synthesize this knowledge. The developer of the NDI published a summary paper in 2008 summarizing a 17-year his-

- **STUDY DESIGN:** Systematic review of clinical measurement.
- **OBJECTIVE:** To find and synthesize evidence on the psychometric properties and usefulness of the neck disability index (NDI).
- **BACKGROUND:** The NDI is the most commonly used outcome measure for neck pain, and a synthesis of knowledge should provide a deeper understanding of its use and limitations.
- **METHODS AND MEASURES:** Using a standard search strategy (1966 to September 2008) and 4 databases (Medline, CINAHL, Embase, and PsychInfo), a structured search was conducted and supplemented by web and hand searching. In total, 37 published primary studies, 3 reviews, and 1 in-press paper were analyzed. Pairs of raters conducted data extraction and critical appraisal using structured tools. Ranking of quality and descriptive synthesis were performed.
- **RESULTS:** Horizon estimation suggested the potential for 1 missed paper. The agreement between raters on quality assessments was high ($\kappa = 0.82$). Half of the studies reached a quality level greater than 70%. Failures to report clear psychometric objectives/hypotheses or to rationalize the sample size were the most common design flaws. Studies often focused on less clinically applicable properties, like construct validity or group reliability, than transferable data, like known

group differences or absolute reliability (standard error of measurement [SEM] or minimum detectable change [MDC]). Most studies suggest that the NDI has acceptable reliability, although intraclass correlation coefficients (ICCs) range from 0.50 to 0.98. Longer test intervals and the definition of stable can influence reliability estimates. A number of high-quality published (Korean, Dutch, Spanish, French, Brazilian Portuguese) and commercially supported translations are available. The NDI is considered a 1-dimensional measure that can be interpreted as an interval scale. Some studies question these assumptions. The MDC is around 5/50 for uncomplicated neck pain and up to 10/50 for cervical radiculopathy. The reported clinically important difference (CID) is inconsistent across different studies ranging from 5/50 to 19/50. The NDI is strongly correlated (>0.70) to a number of similar indices and moderately related to both physical and mental aspects of general health.

- **CONCLUSION:** The NDI has sufficient support and usefulness to retain its current status as the most commonly used self-report measure for neck pain. More studies of CID in different clinical populations and the relationship to subjective/work/function categories are required. *J Orthop Sports Phys Ther* 2009;39(5):400-417. doi:10.2519/jospt.2009.2930
- **KEY WORDS:** cervical spine, outcome measure, reliability, validity

¹Co-director, Hand and Upper Limb Centre Clinical Research Laboratory, St Joseph's Health Centre, London, ON, Canada; Associate Professor, School of Rehabilitation Science, McMaster University, Hamilton, ON, Canada; Funded by a New Investigator Award, Canadian Institutes of Health Research. ²Lecturer and PhD Candidate, School of Physical Therapy, The University of Western Ontario, London, ON, Canada; Funded by a Doctoral Fellowship, Canadian Institutes of Health Research. ³Physiotherapy Student, School of Physical Therapy, The University of Western Ontario, London, ON, Canada. ⁴Emeritus Professor, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada. Address correspondence to Dr Joy MacDermid, Co-director, Hand and Upper Limb Centre Clinical Research Lab, Monsignor Roney Ambulatory Care Centre, 930 Richmond Street, London, ON, N6A 3J4 Canada. E-mail: jmacderm@uwo.ca

tory with the NDI.⁴⁶ The author stated that “The Neck Disability Index has been used in over 300 publications, translated into 22 languages...and endorsed for use by a number of clinical practice guideline committees...making it the most widely used and most strongly validated instrument for assessing suffering disability in patients with neck pain.”

While previous reviews have claimed to be systematic, none have included a key element of systematic review, critical appraisal of the quality of individual studies. This may reflect inherent difficulties in performing the critical appraisal due to lack of instrumentation. The lead author of this current review has published a scale and interpretation guide^{25,26} for this purpose. Psychometric studies are important to establish measurement properties like the relative difficulty of items, appropriate grouping of items into subscales, reliability, validity, and responsiveness. In addition to psychometric properties, clinicians are concerned with issues on feasibility, floor/ceiling effects, availability of different language/cultural adaptations, and administration burden for themselves and their patients. Terms like “clinician/patient friendliness,” or “clinical utility/applicability,” or, as we prefer, “usefulness” have been used in reference to these practical considerations. When therapists try to integrate the NDI into their clinical practice, they are concerned with the psychometric issues but also need information on usefulness. The purpose of this study was to conduct a systematic review that would summarize the quality and content of current research regarding the psychometric properties and usefulness of the NDI.

METHODS

Development of the NDI

VERNON AND MIOR⁴⁷ DEVELOPED the NDI using the Oswestry Low Back Pain Index (OLBPI) as a template for identifying items and a scoring metric. They initially selected 6 items

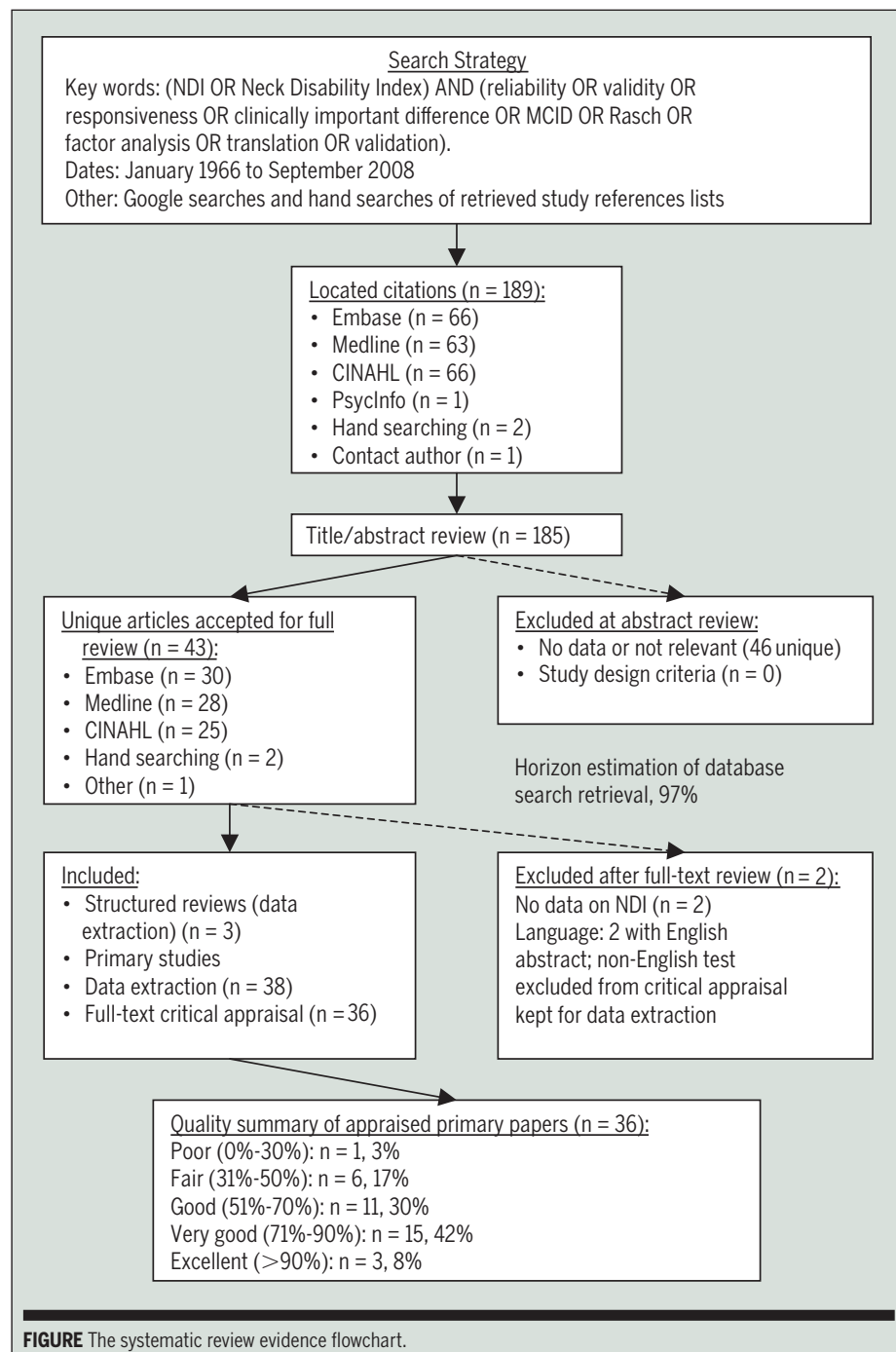


FIGURE The systematic review evidence flowchart.

from the original scale: pain intensity, personal care, lifting, sleep, driving, and sex life, and submitted this to a consulting team. The consulting team added 4 items: headache, concentration, reading, and work, resulting in the 10-item scale. These 10 items were modified for clarity and relevance based on feedback from 5 individuals with a history of whiplash injury and the review team. In the end, the

NDI contained 5 items from the OLBPI, 2 of which were modified, and 5 new items.⁴⁷ The questions are measured on a 6-point scale from 0 (no disability) to 5 (full disability). The numeric response for each item is summed for a score varying from 0 to 50.⁴⁷ Some evaluators choose to provide a score out of 100%, particularly as a strategy to deal with questions left unanswered (**APPENDIX A**).

Literature Search and Study Identification

A database search was performed using Medline, PsychInfo, CINAHL, and Embase, which included papers written between January 1966 and September 2008 (FIGURE). The following keywords were used to search all databases for eligible studies: (NDI OR Neck Disability Index) AND (reliability OR validity OR responsiveness OR clinically important difference OR MCID OR Rasch OR factor analysis OR translation OR validation). In addition, we conducted Google searches and hand searches of retrieved study references lists. The number of studies retrieved and the net results of abstract/title and full review are noted in the FIGURE. We conducted a Horizon estimation to evaluate the potential that any articles were missed and the efficiency of our search using a procedure developed by Foster and Goldsmith for SAS software (full details available from author Charlie Goldsmith).¹⁴

The first stage of study identification was of title/abstracts, which were independently reviewed by at least 2 of the authors. An article was accepted if it met the following inclusion criteria: reported on at least 1 psychometric property of the NDI in patients with neck pain and was written in French or English (2 with English abstracts and non-English full text [1 Dutch, 1 Spanish] were included for abstract data extraction but not full critical appraisal). In total, 37 primary studies, 3 structured reviews, and 1 unpublished in press manuscript proceeded to a full review (FIGURE, TABLE 1).

The data extraction and review process was conducted by pairs of raters and was based on the use of structured instruments and a predetermined consensus process.^{25,26} All authors met for a calibration review, in which they independently reviewed 1 paper then met and discussed each item to clarify the meaning and interpretation of critical appraisal items. Following this, pairs of raters independently evaluated an assigned subset of articles using previously developed data

extraction forms and quality appraisal tools. The data extraction form was developed by adaptation of categories used in a previous psychometric review by Euchaute et al,¹³ with additions from the first author, who also developed an accompanying guide (APPENDIX B, AVAILABLE ONLINE). Critical appraisal forms and interpretation guide used to assess individual study quality were developed by the lead author.^{25,26} The items from the quality tool are listed in TABLE 2, with the associated ratings for individual studies. Interrater reliability on quality ratings was calculated based on preconsensus scores of rater pairs; the overall estimated kappa was 0.82, with agreement on individual items varying from 0.43 to 1.00. We used the following consensus process when independent assessments differed:

1. First, raters clarified if the disagreement was based on factual content or the extent of compliance with the item. Factual content/oversights (the most common source of disagreement) were resolved by recheck of the original manuscript.
2. Pairs of raters consulted the first author/instrument developer (J.M.) for clarification of the meaning or interpretation of specific items of the scale if they felt this was contributing to their disagreement.
3. Raters then discussed their perceptions of the extent to which the study manuscript complied with item requirements (if something was not documented in the manuscript it was treated as “not done”).
4. At this point, if the raters continued to disagree on an item by 1 point, the raters decided if they were comfortable with the default compromise of assigning the lesser of the 2 scores. If the raters were not comfortable with this resolution or continued to disagree by 2 points, an adjudicator was required, although this was not required during the study.

Each paper's score was converted into a percentage because 1 item was based on follow-up and some psychometric stud-

ies are cross-sectional, leaving unequal denominators for different studies. We rank ordered studies on quality and considered this ranking when making conclusions and recommendations, although there was no formal mechanism to weight conclusions, based on the quality of the associated source document.

RESULTS

THE HORIZON ANALYSIS¹⁴ INDICATED a possibility that our search strategy missed 1 article (95% confidence interval, 0-2; 97% yield). An optimal search strategy would have been Embase first, then CINAHL, and then Medline. PsychInfo did not contribute any information after these 3 databases.

In total, 41 studies were identified that addressed at least 1 psychometric property of the NDI (TABLE 1). Two neck disability instrument reviews were identified, although neither included formal critical appraisal.^{34,37} Similarly, a recent comprehensive review of the NDI by the developer did not include formal critical appraisal of individual studies.⁴⁶ The primary studies crossed different populations, interventions, and time intervals, and addressed different psychometric properties. Quality of the individual studies was variable, ranging from 21% to 96%, with 37% of papers reaching or exceeding a score of 75% on the quality rating (TABLE 2). The most common flaws observed in the psychometric articles were (1) not reporting specific psychometric hypothesis/objectives, (2) inadequate sample size calculations/justification, and (3) absence of error estimates such as confidence intervals or standard error of measurement (SEM). A descriptive synthesis of the findings for psychometric properties across all identified studies is summarized in TABLES 3 through 5. Due to the heterogeneity of study populations and properties evaluated, no meta-analyses were performed. Most studies addressed a spectrum of psychometric properties, but few were comprehensive. Few studies provided specific conclusions or

TABLE 1

SUMMARY OF STUDIES ADDRESSING PSYCHOMETRICS OF THE NDI

Study	Population	n	Properties Evaluated	Intervention/Retest Interval
Ackelman et al 2002 ¹	<u>Patients:</u> 59 Swedish patients (28 M, 31 F); twenty patients were in the acute phase after a neck sprain, 19 had chronic NP, and 20 had no NP but had other musculoskeletal symptoms	59 (38 for modified version)	Reliability, validity, responsiveness	<u>Intervention:</u> PT <u>Retest interval:</u> 5 occasions, retest within 48 hours, at 3 w and 3 mo
Andrade et al 2008 ²	<u>Patients:</u> nonspecific or posttraumatic NP <u>Site:</u> Spain	48	Spanish translation, reliability, responsiveness	<u>Intervention:</u> ND
Aslan et al 2008 ³	<u>Patients:</u> NP <u>Included:</u> patients with at least 3 mo of NP	88	Reliability, construct validity	<u>Intervention:</u> PT <u>Retest interval:</u> 7 d
Bolton 2004 ⁵	<u>Patients:</u> nonspecific NP <u>Site:</u> 8 chiropractic clinics and 1 teaching clinic	125	Responsiveness	<u>Intervention:</u> chiropractic <u>Retest interval:</u> pre-Rx and 4-6 wk postinitial Rx
Chan et al 2008 ⁷	<u>Included:</u> patients with chronic nontraumatic NP with a duration of greater than 3 mo <u>Excluded:</u> cervical radiculopathy, ankylosing spondylitis, rheumatoid arthritis, neck trauma	20	Construct validity	<u>Intervention:</u> ND <u>Retest interval:</u> Cross-sectional
Chok et al 2000 ⁸	<u>Patients:</u> NP <u>Site:</u> an outpatient setting <u>Included:</u> 2 cohorts, 22 tested twice for reliability and 24 that completed baseline and discharge assessments	46	Reliability, validity	<u>Intervention:</u> PT <u>Retest interval:</u> admission and discharge
Cleland et al 2006 ¹⁰	<u>Patients:</u> 47% F; mean age, 51 y; symptom duration, 14 wk <u>Site:</u> PT Clinics	38	Reliability, validity, responsiveness	<u>Intervention:</u> PT <u>Retest interval:</u> baseline evaluation and after 5-7 visits
Cleland et al 2008 ⁹	<u>Patients:</u> stable; 60% F; mean age, 42 y; symptom duration, 83 d; improved, 25% F; mean age, 43 y; symptom duration, 46 d <u>Included:</u> age 18-60 y, NDI score greater than 10%, NP with or without referral of symptoms to the upper extremity or extremities. <u>Excluded:</u> any signs or symptoms consistent with a nonmusculoskeletal etiology, WAD within the past 6 wk, central nervous system involvement, 2 or more signs consistent with nerve root compression, prior surgery to the cervical or thoracic spine, or pending legal action regarding their NP	138	Reliability (SEM and ICC), construct validity, responsiveness (SRM and MDC)	<u>Intervention:</u> manual therapy <u>Retest interval:</u> after 1 session
Cook et al 2006 ¹¹	<u>Patients:</u> 62% M; aged, 18-89 y (mean, 43 y) <u>Site:</u> midsize hospital in southern Brazil	203	Brazilian Portuguese translation, validity, responsiveness	<u>Intervention:</u> ND <u>Retest interval:</u> 10 subjects retested randomly at 24 h and a separate 10 subjects retested at 7 d
Gay et al 2007 ¹⁵	<u>Patients:</u> 7 M, 16 F; mean age, 50 y; chronic uncomplicated NP	23	Primary purpose to evaluate another scale; NDI: internal consistency, responsiveness	<u>Intervention:</u> 1 of 3 manual therapy approaches, plus heat, home exercise program <u>Retest interval:</u> 4 wk
Goolkasian et al 2002 ¹⁶	<u>Patients:</u> 12 M, 21 F; mean age, 43 y; chronic NP <u>Site:</u> PT clinics <u>Included:</u> received PT that they considered inadequate <u>Excluded:</u> surgical candidates	33	Responsiveness	<u>Intervention:</u> group 1, PT; group 2, involved in a study of botulin toxin for NP <u>Retest interval:</u> within 1 wk
Hains et al 1998 ¹⁷	<u>Patients:</u> 62% F; mean age, 39 y; many occupations; acute, subacute, and chronic NP <u>Site:</u> chiropractic college clinics in Vancouver and Toronto <u>Included:</u> English speaking, NP, 17 y or older	237	Validity	<u>Intervention:</u> unknown <u>Retest interval:</u> unknown
Heijmans et al 2002 ¹⁸	<u>Patients:</u> chronic WAD	61	Dutch translation, reliability, validity	Unknown
Hoving et al 2003 ¹⁹	<u>Patients:</u> mean age, 40 y; 59% of patients were M; all patients were less than 24 mo since MVA; varying severity and duration of WAD (level of WAD: 23% I, 42% II, and 6% III), n = 71 <u>Site:</u> patients recruited from PT clinics, GPs, and community rheumatology clinic into private PT practice	71	Validity	<u>Intervention:</u> none

[LITERATURE REVIEW]

TABLE 1

SUMMARY OF STUDIES ADDRESSING PSYCHOMETRICS OF THE NDI (CONTINUED)

Study	Population	n	Properties Evaluated	Intervention/Retest Interval
Humphreys et al 2002 ²⁰	<u>Patients:</u> NA <u>Site:</u> 8 chiropractic clinics and 1 teaching clinic	125	Reliability, validity, responsiveness	<u>Intervention:</u> chiropractic <u>Retest interval:</u> postinitial Rx, 4-6 d post-Rx
Kose et al 2007 ²¹	<u>Patients:</u> 83% F; age, 18-55 y; >6 wk of NP <u>Excluded:</u> inflammatory arthritis, spine injury, comorbidity, infection	102	Reliability, responsiveness	<u>Intervention:</u> PT 5/wk for 3 wk <u>Retest interval:</u> 1-3 d without therapy
Kovacs et al 2008 ²²	<u>Patients:</u> acute, subacute, and chronic who visited their physician for NP <u>Site:</u> 9 primary care and 12 specialty services from 9 regions in Spain <u>Excluded:</u> functional illiteracy, neck trauma, surgery, systemic disease, central nervous system	221 (54 from pilot, plus 167)	Spanish translation, reliability, validity, responsiveness	<u>Intervention:</u> ND <u>Retest interval:</u> 1-d retest and re-evaluation at 14 d
Kumbhare et al 2005 ²³	<u>Patients:</u> the first group comprised of 81 patients from Ontario Canada who had grade II WAD; second group subjects from a convenience sample who did not have NP and who were not being treated <u>Site:</u> tertiary outpatient clinic	241 (81 patients, 160 controls)	Reliability, validity, responsiveness	<u>Intervention:</u> PT <u>Retest interval:</u> follow-up varied between 4-12 w
Lee et al 2006 ²⁴	<u>Patients:</u> 17 M, 31 F; age, 18-69 y, with a primary diagnosis of NP <u>Site:</u> patients were chosen from 3 hospitals and 7 clinics in 5 Korean cities	180	Reliability, validity, responsiveness	<u>Intervention:</u> ND <u>Retest interval:</u> first and last visit or 7th or 8th Rx
McCarthy et al 2007 ²⁹	<u>Patients:</u> NP <u>Site:</u> spinal clinic	160 (34 retest)	Reliability, validity	<u>Intervention:</u> various <u>Retest interval:</u> 1-2 wk
Miettinen et al 2004 ³⁰	<u>Patients:</u> NP who had been involved in an MVA <u>Site:</u> Finland	144	Validity, responsiveness	<u>Intervention:</u> ND <u>Retest interval:</u> 1 and 3 y post-MVA
Mousavi et al 2007 ³¹	<u>Patients:</u> NP >3 mo 30%, >1 year 33%, gradual onset 78%; 54% F; baseline VAS, 55/100 <u>Site:</u> primary care or PT clinics	185	Iranian translation, relative reliability, internal consistency, floor/ceiling, convergent validity, interpretability	<u>Intervention:</u> ND <u>Retest interval:</u> none
Nieto et al 2008 ³³	<u>Patients:</u> Catalan-speaking, with subacute pain of less than 3 mo following whiplash injury from 10 different practices; aged 18-65 y <u>Excluded:</u> fracture dislocation, prior history of chronic pain, neurological injury, 2 or more missing items on NDI	150	Factor validity, translation	<u>Intervention:</u> routine Rx
Pietrobon et al 2002 ²⁴	Structured review of different neck disability scales	NA	Validity	NA
Pool et al 2007 ³⁵	<u>Patients:</u> 61% F; mean age, 46 y; nonspecific NP; duration 2-6 wk, 48%; 7-12 wk, 26%; >13 wk, 26%; NDI, 14; pain, 6/10	183	CID	<u>Intervention:</u> random allocation to PT, manual therapy, GP care
Rebbeck et al 2007 ³⁶	<u>Patients:</u> 3 cohorts primary care insurance (early and late); employed: 80%, 68%, 66%; F: 80%, 70%, 67%; NDI score, 23; NR, 18	(99, 134, 250)	Convergent validity, responsiveness	ND
Resnick 2005 ³⁷	An overview of instruments in a structured review by a single author and evaluator; include a content analysis at item level	NA	Validity	NA
Riddle et al 1998 ³⁸	<u>Patients:</u> 64% F; 88% of patients 50 years or younger; variety of neck disorders; most common neck strain or NP <u>Site:</u> 1 of 4 clinics in Richmond, VA <u>Excluded:</u> under treatment for problems other than c-spine	146 (69 for retest)	Validity, responsiveness	<u>Intervention:</u> unknown <u>Retest interval:</u> admission and discharge
Scolasky et al 2007 ³⁹	<u>Patients:</u> NP cohort cervical radiculopathy or myelopathy, adequate mental capacity within 4 wk of surgery <u>Sites:</u> 23 clinics <u>Excluded:</u> revisions/tumors	534	Convergent validity	<u>Intervention:</u> anterior cervical decompression and fusion

TABLE 1

SUMMARY OF STUDIES ADDRESSING PSYCHOMETRICS OF THE NDI (CONTINUED)

Study	Population	n	Properties Evaluated	Intervention/Retest Interval
Stewart et al 2007 ⁴²	<u>Patients:</u> 17 M, 31 F; chronic WAD (>3 mo); "mildly" disabled preinjury and presented for medical care within 1 mo of MVA; a score of at least 20% on primary outcome measures: NPRS, pain bother-someness, PSFS <u>Excluded:</u> previous neck surgery, known or suspected red flags, nerve root compromise, contraindicated to exercise, severe or greater depressive symptoms, no neck radiograph since MVA, current PT neck Rx, poor English	134	Reliability, responsiveness	<u>Intervention:</u> 6-wk individualized submaximal exercise program, supervised by PT <u>Retest interval:</u> 3 sessions in week 1, 2; 2 sessions week 3, 4; 1 session weeks 5, 6; 2 follow-up phone calls
Stratford et al 1999 ⁴³	<u>Patients:</u> NP <u>Site:</u> 5 PT clinics in North America	50	Reliability, validity, responsiveness	<u>Intervention:</u> PT <u>Retest interval:</u> 1-3 wk
Trouli et al 2008 ⁴⁴	<u>Patients:</u> NP <u>Site:</u> primary care in 3 rural centers <u>Excluded:</u> symptoms below the elbow, neurological signs, positive upper limb tension test, 6 cases removed because of 2 missing items on NDI	65 (43 classified as stable)	Factor validity, content validity, internal consistency, retest reliability, translation	<u>Intervention:</u> routine Rx
van der Velde et al ⁴⁵	<u>Patients:</u> 55% or more F in 2 cohorts from previous studies of n = 188 and the other n = 333 <u>Included:</u> pain for 5-6 wk, baseline pain of 5-6/10, and total NDI of 13-14 <u>Site:</u> different clinics	521	Factor analysis, item consistency, Rasch modeling of item, evaluation of reduced 8-item scale and differential item functioning of items	<u>Intervention:</u> ND
Vernon 2000 ⁴⁸	Combination of instrument review (NDI and other), plus some impairment sincerity of effort testing	NA	Content validity (comparison of items)	NA
Vernon 2008 ⁴⁶	A structured review by a single author (the NDI developer) that includes psychometrics, prediction and effect sizes across treatment RCTs that used the NDI; no formal critical appraisal of studies	NA	NA	NA
Vernon et al 1991 ⁴⁷	<u>Patients:</u> 17 M, 31 F; 70% WAD in past 4-6 wk; 30% chronic nontraumatic <u>Site:</u> Clinic of Canadian Memorial Chiropractic College	48	Reliability, validity	<u>Intervention:</u> none <u>Retest interval:</u> 2 d
Vos et al 2006 ⁴⁹	<u>Patients:</u> 119 F, 68 M; recurrent acute NP lasting no longer than 6 wk, with pain-free interval of at least 3 mo <u>Site:</u> referred by GPs in Rotterdam and region <u>Inclusion criteria:</u> aged 18 and sufficient Dutch <u>Exclusion criteria:</u> specific cause of NP (ie, vascular or neurological disorders, etc)	187	Reliability, responsiveness	<u>Intervention:</u> none <u>Retest interval:</u> 1-wk follow-up
Westaway et al 1998 ⁵⁰	<u>Patients:</u> Age, 12-80 y; nonchronic NP of musculoskeletal origin "cervical dysfunction" <u>Site:</u> referred by physician to PT	31	Reliability, validity, responsiveness	<u>Intervention:</u> routine Rx <u>Retest interval:</u> initial assessment, 72 h, and following 1-4 w of Rx by 1 certified Canadian orthopaedic manipulative PT
White et al 2004 ⁵¹	<u>Patients:</u> chronic mechanical NP <u>Site:</u> recruited from PT and rheumatology waiting lists at 2 hospitals in UK <u>Excluded:</u> undergoing litigation, systemic or other major disorder	133	Validity	<u>Intervention:</u> in treatment RCT <u>Retest interval:</u> before Rx initiated, at 1 wk and 8 wk
Wlodyka-Demaille et al 2002 ⁵³	<u>Patients:</u> mean age, 49 y; NP <u>Site:</u> patients recruited from outpatient clinics, rehabilitation department, rheumatology department (ie, tertiary care teaching hospital and outpatient clinic)	101	Reliability, validity	<u>Intervention:</u> none <u>Retest interval:</u> 24 h
Wlodyka-Demaille et al 2004 ⁵²	<u>Patients:</u> 43 F, 28 M; mean age, 49 y <u>Site:</u> patients recruited from outpatient clinics, rehab department, rheumatology department (ie, tertiary care teaching hospital and outpatient clinic)	71	Validity, responsiveness	<u>Intervention:</u> ND <u>Retest interval:</u> 11 ± 2 mo

Abbreviations: CID, clinically important difference; F, females; GP, general practitioner; ICC, intraclass correlation coefficient; LBP, low blood pressure; M, males; MDC, minimal detectable change; MVA, motor vehicle accident; ND, not described; NDI, Neck Disability Index; NP, neck pain; NPRS, Numeric Pain Rating Scale; NR, numeric rating; PSFS, Pain-Specific Functional Scale; PT, physiotherapy; RCT, randomized controlled trial; Rx, treatment; SEM, standard error of measurement; SRM, standardized response mean; UK, United Kingdom; VA, Virginia; VAS, visual analog scale; WAD, whiplash-associated disorder.

[LITERATURE REVIEW]

numbers that could readily be integrated into clinical practice but, rather, tended to rely on generalities when making conclusions. The type of data collected during NDI validation studies was typically comprised of less clinically useful data,

TABLE 2

QUALITY OF STUDIES ON THE PSYCHOMETRIC OF THE NECK DISABILITY INDEX

Study [†]	Item Evaluation Criteria*												Total (%)
	1	2	3	4	5	6	7	8	9	10	11	12	
Cleland et al 2008 ⁹	2	2	1	2	2	2	2	2	2	2	2	2	96
Van der Velde et al ⁴⁵	2	2	2	1	2	2	2	2	2	2	2	2	96
Kumbhare et al 2005 ²³	2	2	2	2	2	2	2	2	1	2	2	2	96
Westway et al 1998 ⁵⁰	2	2	2	2	0	2	2	1	2	2	2	2	88
Cook et al 2006 ⁴¹	2	2	2	2	0	2	1	2	2	2	2	2	88
Cleland et al 2006 ¹⁰	2	2	2	2	0	2	1	2	2	2	2	2	88
Trouli et al 2008	1	2	1	2	1	2	2	1	2	2	2	2	83
Hoving et al 2003 ¹⁹	2	2	1	2	0	NA	2	2	2	2	1	2	82
Aslan et al 2008 ³	2	2	1	2	0	1	2	2	2	2	1	2	79
Stratford et al 1999 ⁴³	2	1	1	1	1	2	2	1	2	2	2	2	79
McCarthy et al 2007 ²⁹	2	1	1	2	1	1	2	2	1	2	2	2	79
Wlodyka-Demaille et al 2002 ⁵³	1	2	1	2	1	2	1	1	2	2	1	2	75
Riddle et al 1998 ³⁸	2	2	2	2	0	0	2	1	2	2	1	2	75
Lee et al ²⁴	2	1	2	2	0	0	2	2	1	2	2	2	75
Stewart et al 2007 ⁴²	1	2	2	1	0	NA	2	2	2	2	1	1	73
Pool 2007 ³⁵	2	2	1	0	1	1	1	1	2	2	2	2	71
Kose et al 2007 ²¹	1	2	1	2	1	2	1	1	2	2	1	1	71
Kovacs et al 2008 ²²	1	2	1	2	1	1	1	2	1	2	2	1	71
Hains et al 1998 ¹⁷	2	1	2	1	0	NA	2	1	2	1	1	2	68
Ackelman et al 2002 ¹	1	2	1	2	0	2	1	2	2	2	0	1	67
Vos et al 2006 ⁴⁹	2	2	1	1	0	0	2	2	1	2	2	1	66
Vernon et al 1991 ⁴⁷	2	1	1	2	0	2	1	1	2	2	0	2	66
Wlododycka et al ⁵²	2	2	0	0	0	1	1	1	2	2	2	2	63
White et al ⁵¹	1	2	1	1	0	1	1	1	2	2	1	2	63
Goolkasian et al 2002 ¹⁶	1	1	2	1	0	1	1	1	2	1	2	2	63
Gay et al 2007 ¹⁵	2	2	1	2	0	1	1	0	1	2	2	1	63
Nieto et al 2008 ³³	2	2	1	1	1	NA	1	1	2	2	0	1	61
Pietrobon et al 2002 ³⁴	2	0	1	2	1	0	2	NA	2	1	0	2	59
Mousavi et al 2007 ³¹	2	1	1	2	1	0	1	0	1	2	1	1	59
Scolasky 2007 ³⁹	1	2	0	0	2	-	2	2	1	1	0	0	50
Bolton 2004 ⁵	1	1	0	1	0	0	2	1	2	1	1	2	50
Chan et al 2008 ⁷	1	2	0	0	0	NA	1	1	2	2	0	1	43
Humphreys et al 2002 ²⁰	1	1	1	1	0	0	0	0	2	2	0	2	42
Rebbeck 2007 ³⁶	0	1	1	1	2	-	0	0	1	1	1	0	36
Chok et al 2005 ⁸	1	1	0	1	0	0	1	1	1	0	1	1	33
Miettenen et al 2004 ³⁰	0	2	0	0	0	0	0	0	1	0	1	1	21

Abbreviations: NA, not applicable to paper.

* Evaluation criteria: 1. Thorough literature review to define the research question; 2. Specific inclusion/exclusion criteria; 3. Specific hypotheses; 4. Appropriate scope of psychometric properties; 5. Sample size; 6. Follow-up; 7. The authors referenced specific procedures for administration, scoring, and interpretation of procedures; 8. Measurement techniques were standardized; 9. Data were presented for each hypothesis; 10. Appropriate statistics-point estimates; 11. Appropriate statistical error estimates; 12. Valid conclusions and clinical recommendations.

[†] Papers where NDI evaluations were performed while evaluating a different measure as the primary purpose. Quality scores were rated in content for the NDI-related methods. Three structured reviews underwent data extraction but no critical appraisal.^{34,37,46} Two papers had data extraction from the English abstract and study tables but no critical appraisal, because they were in non-English text.^{2,18} One paper had item comparison but no NDI data, so no full critical appraisal.⁴⁶

like correlations indicating construct convergent validity, versus more useful information, like known group differences that could be used as comparative data for clinical comparisons. Similarly, group reliability, such as intraclass correlation coefficients (ICCs), was reported more often than more useful indicators of absolute measurement error (like SEMs, mean retest differences, or minimal detectable change [MDC]).

Summary of Previous Structured Reviews

Three papers were identified where the authors performed a comprehensive structured review with use of a formal search for articles, but without use of multiple raters or formal critical appraisal. The first review³⁴ is important because it established the relative strengths and weaknesses of the NDI when compared to other scales. This review was based on a formal search of 2 databases, MEDLINE and CINAHL, citation tracking using the citation index, hand searching of relevant journals, and correspondence with experts to find additional papers (up to year 2000). No formal quality evaluation was performed. Five standardized neck pain scales were identified and compared qualitatively in terms of content and measurement properties. These authors concluded that 3 of the scales were similar in terms of structure and psychometric properties: the NDI, the Copenhagen Neck Functional Disability Scale, and the Northwick Park Scale. It was suggested that the NDI had the strength of being most studied amongst these 3 and was at that time the only instrument revalidated in different study populations. The authors of this review suggested other scales had important limitations. These included that the Neck Pain and Disability Scale must be read to the patient and the Patient-Specific Functional Scale was considered “very sensitive to functional changes in individual patients, but comparisons between patients are virtually impossible.”³⁴

TABLE 3		SUMMARY OF RELIABILITY PROPERTIES	
Reliability	Data Extracted		
Mean retest difference	As per Altman and Bland technique, where reliability is assessed as mean difference between retests and the 2 standard deviation brackets around that difference, -1.5 ± 3^{44}		
SEM	Patients classified as stable, using GRC scores of 3 to $-3/7$; SEM, 0.64^{44} Patients classified as stable over 1 wk (GRC, 3 to $-3/7$); SEM, 8.4^9		
MDC	Patients classified as stable (GRC 3 to $-3/7$); MDC, 19.6^9 MDC, 10.5 in patients with neck pain ³⁵ MDC, 10.2 in patients with cervical radiculopathy ¹⁰ MDC, 5 with 90% CI in patients with neck pain ⁴⁸ MDC, 4.7 with 90% CI ⁴³ Patients classified as stable over 1 wk (GRC, 3 to $-3/7$); MDC, 1.78^{44} MDC, 1.66 in WAD ⁴⁹		
Test-retest reliability coefficients	1 d, $r = 0.92$ (chronic) to 0.93 (acute) ¹¹ 1 d, 0.92^{11} 1 d, 0.93^{53} 2 d, $r = 0.94-0.99$ (chronic) to $0.81-0.89$ (acute) ¹ 2 d, 0.97^1 3 d, patients in acute condition, $r = 0.73^{50}$ 1-3 d, in patients with >6 wk of neck pain 0.86^{21} 3-7 d, in healthy controls, 0.90^{24} 7 d, $r = 0.48$ (chronic) to 0.90 (acute) ¹¹ 7 d, 0.90^{49} 7 d, 0.93 7 d, 0.98^3 7-14 d, 0.96 7-14 d, -0.93^{29} 7-21 d, $r = 0.94^{43}$ 21 d, 0.95^1 90 d, 0.94^1 0.98^{22} 0.97 ($n = 30/185$) ³¹ Patients classified as stable on GRC, ICC = 0.50^9 $k = 0.90$ in first cohort ⁸		
Abbreviations: CI, confidence interval; GRC, global rating of change; ICC, intraclass correlation coefficient; MDC, minimal detectable change; SEM, standard error of the measurement; WAD, whiplash-associated disorder.			

The results of a subsequent systematic review of studies on outcome measures for the cervical spine published up until 2004 suggested that instrument development is ongoing as it identified an increased number of neck pain scales (a total of 11 scales). Again, while databases were used to identify papers, other elements of systematic review were not performed, including use of multiple raters/data extractors, quality ratings for indi-

vidual studies, or a formal process to synthesize results. Eight English-language (specific-to-neck) scales and 3 (nonspecific validated for the neck) scales were reviewed. The instruments (and year published) are listed in **TABLE 6**. The psychometric properties of each scale were reviewed in a qualitative manner, while the content of instruments was compared more quantitatively in a summary table. This item analysis indicated that the most

[LITERATURE REVIEW]

TABLE 4

SUMMARY OF VALIDITY PROPERTIES

Validity	Data Extracted
Content (including analyses of question, appropriateness, missing items, ceiling/floor effects)	<ul style="list-style-type: none"> • Comparison of items across NDI, NPQ, and the Copenhagen Neck Functional Disability Index; items on all 3: sleeping and reading; items on 2 scales: pain, personal care, lifting, concentration, work, driving, recreation⁴⁸ • Authors used a patient interview to elicit problems called the problem elicitation technique to identify functional problems in patients with chronic nontraumatic neck pain. It was determined that $\geq 75\%$ of patients identify problems with sleep, mobility, role activity, emotion, and symptoms. Sleep disturbance had the highest prevalence. More than half of subjects identified difficulties with frustration, driving, and lifting. Less common but mentioned by more than 30% of patients were looking into cupboards, gardening, headaches, housework, working overhead, and general exercise. The highest mean severity scores were found in depression, cooking, and sitting upright. Individual problems ranked most important by subjects were driving, sleep disturbance, and frustration. Of the 11 problems identified by most subjects, 6 were included on the NDI.⁷ • NDI was not sensitive to the items concentration (1.4, 1.1) or personal care (0.7, 0.8).¹⁹ • NDI did not include the extremes of very easy or very difficult items, many were similar in difficulty and therefore potentially redundant.²⁴ • Social consequences are an important part of subject's situation that is not mentioned in NDI.¹ • Validated for multiple origins of neck pain, function, and clinical signs and symptoms.³⁴ • 8% of items missing, driving most often left missing due to not applicable.²¹ • In Iran, 42% did not answer driving, 9% reading.³¹ • Looked for ceiling effects but not detected.³¹ • 8.5% of patients had initial scores within 1 MDC distance from the best possible answer (0) revealing no ceiling effects according to a 15% criterion.⁴⁴ • 10/50 asked for clarification during Spanish adaptation.²² • Most commonly missing items (on baseline assessment and 1 wk later) included lifting 11%, reading 9%, driving 45%, recreation 1%⁴⁴
Internal consistency	<ul style="list-style-type: none"> • Consistently high Cronbach alpha: 0.70-0.96.^{11,15,17,24,29,43,47}
Factor validity	<p><u>Support for 1 factor</u></p> <ul style="list-style-type: none"> • 1 factor confirmed 84% variance on first factor^{11,17} • 1 factor² <p><u>Support for 2 factors</u></p> <ul style="list-style-type: none"> • Factor analysis: 2 factors extracted (eigenvalues, >1.0)⁵³ • Factor 1, function and disability (items: 2, 6-8, 10)⁵³ • Factor 2, neck pain (items: 1, 4, 5, 9)⁵³ • Factor 1, pain and interference with cognitive functioning³³ • Factor 2, functional disability • Using an oblique rotation technique, 1 factor "pain and interference of cognitive functioning," explained 42% of the variance and contained the following items: pain intensity, reading, headaches, concentration, and sleeping. The second factor on functional disability included the items personal care, lifting, work, driving, and recreation. The correlation between these factors was 0.66. Because of previous reports of a 1-factor solution, analyses were completed to determine differences between the 1- and 2-factor solutions, the latter being significantly better. A scree plot confirmed a 2-factor solution as optimal.
Construct, convergent	<p><u>Correlations with other scales reported >0.70</u></p> <ul style="list-style-type: none"> • PSFS,⁵⁰ NPQ, NPDS, DRI, VAS (activity-chronic), VAS (pain and activity, acute).^{119,50,53} VAS pain,²¹ VAS pain, CSQ, NPQ²² • VAS bodily pain³¹ <p><u>Correlation with other scales reported as moderate (0.30-0.70):</u></p> <ul style="list-style-type: none"> • SF-36 PCS, SF-36 MCS, HADS, NPAD, VAS (pain, chronic).^{116,19,38,53} • VAS disability, PCSS, MCSS, NPQ baseline, CSQ baseline • VAS disability, physical rating, muscle spasm, neck ROM, neck sensitivity²¹ • General health, vitality, social function, physical function³¹ • COS: pain in neck, arms, physical symptoms, function. Psychological distress, healthcare use,³⁹ a Patient-Specific Problem Elicitation Technique⁷; r for VAS test and retest groups of 0.51 and 0.62³
Construct, known group differences	<p><u>Detected differences between</u></p> <ul style="list-style-type: none"> • Disability levels: no disability, 18; mild, 27; moderate, 40; severe, 44 and very severe, 70. • Pain: no pain, 22; mild, 30; moderate, 36; severe, 6 and very severe, 61²² • Amount of change: 10.5, completely recovered; much improved, 8; slightly improved, 3.7; no change, 0.26; important change, 9³⁵ • Recovered, <8; mild disability, 10-28; moderate/severe disability, $>30$⁴¹ • At 24 wk, recovered, 14; persistent pain, 28^{32,46}

Abbreviations: CSQ, Cervical Spine Outcomes Questionnaire; DRI, Disability Rating Index; HADS, Hospital Anxiety and Depression Scale; MCSS, Multi-country Survey Study; NPDS, Neck Pain and Disability Score; NPAD, Neck Pain and Disability Scale; NPQ, Northwick Park Neck Pain Questionnaire; PCSS, Pain Coping Strategy Scale; PSFS, Patient-specific Functional Scale; ROM, range of motion; SF-36 MCS, Short-Form 36 mental component scale; SF-36 PCS, Short-Form 36 physical component scale; VAS, visual analogue scale.

common content items on neck disability scales were self-care/activities of daily living (ADL), work, standing/running, pain intensity, and driving.³⁷

Most recently, the developer of the NDI published a comprehensive 17-year review of the NDI.⁴⁶ This review includ-

ed a historical perspective, a summary of psychometric properties from 22 different studies, a list of translations available from the MAPI website, a summary of results from prognostic studies that use the NDI, and tables listing mean change scores in effect sizes with different treat-

ments, including manipulation, exercise, mobilization, acupuncture, medication, cervical pillow, laser, and relaxation therapy. This publication provided the most extensive review of the NDI to date but did not include independent critical appraisal.

Summary of Primary Studies

Readability/Language and Cultural Translation A number of papers have addressed issues around readability, usually in the context of language and cultural translations. Wlodyka-Demaille et al⁵³ reported that the concepts of “leisure” and “social activities” had different meanings in French and American cultures. For that reason, they provided examples in these areas to make the questionnaire more understandable for this population.⁵³ Through a series of translating and back translating, Lee et al²⁴ concluded that their translation process retained the sound measurement properties of the original English version in the Korean version. Cook et al¹¹ ensured good readability of the Brazilian/Portuguese version of the NDI through a committee who reviewed the translator reports. They reached consensus on discrepancies in 4 areas, including semantic equivalence, experiential equivalence, idiomatic equivalence, and conceptual equivalence.¹¹ In the Swedish version, Ackelman and Lindgren¹ changed each item of the tool to specify that only disability due to neck pain was of interest. In the Spanish version, 16% of patients had comprehension difficulties, but these did not appear to be related to educational level and reliability was excellent.² Others identified that the driving item had a high rate of nonresponse in an Iranian population.³¹ Overall readability in the English and all translated versions is deemed to be acceptable, although some patients require support to understand the items, and nonresponse is more common for task items like driving and, to a lesser extent, reading.

Not all translations have been published in peer-reviewed literature. The developer reported that he worked with an independent organization (www.proqolid.org) to produce additional translations through standardized translation methodology (although psychometric testing was not performed). These translations included English (Australian/UK/US), Danish, Finnish,

French (Canadian/Switzerland/Germany), Italian, Norwegian, Polish, Spanish (Spain/US).

Administration Burden Few authors specifically address or state how they measured the time taken to complete the NDI. However, all agree that only a short amount of time is required. Stratford et al⁴³ reported that the time for patients to complete the NDI was about 3 minutes, while Wlodyka-Demaille et al⁵³ reported that it took a mean (SD) of 7.4 (6.8) minutes (range, 1-60 minutes). There was only a single report of therapist burden stating that it takes 8-10 minutes to complete and 5 minutes to analyze the (Dutch) NDI.¹⁸ Stratford et al⁴³ commented that the questionnaire could be completed in the waiting room and thus did not add any additional time to the patient’s visit.

Interpretability/Subgrouping of Differential Outcomes

NDI scores vary from 0 to 50, where 0 is considered “no activity limitation” and 50 is considered “complete disability.”⁴⁷ Originally, no explicit recommendations were made by the developers on the handling of missing items, or minimum number of items required for validity. More recently, it has been suggested that if 3 or more items are missing, the score may not be valid.⁴⁶ The majority of authors report the NDI out of a total of 50; however, some authors^{1,9} provide a percentage score as a strategy to account for questions that are left unanswered.^{1,10} In the study by Ackelman and Lindgren,¹ if more than 2 items were left unscored, the subjects were not included.

Vernon and Mior⁴⁷ offered the following interpretation of NDI scores: 0 to 4, no disability; 5 to 14, mild disability; 15 to 24, moderate disability; 25 to 34, severe disability; and greater than 35, complete disability. However, no process was described for how these ratings were derived and no validation of these categories was performed. One group of authors set the “normal limit” of the NDI between 0 and 20 points.³⁰ Again, the methodology behind setting this benchmark was not described, and this cutoff has not been

validated. Sterling et al⁴⁰ used data from clinical studies to define patients who had recovered as having NDI scores of less than 4 (8%), those with mild disability as having scores of 5 to 14 (10%-28%), and those with moderate to severe disability as having scores of greater than 15 (30%). Conversely, Nederland³² found that patients defined as recovered at 24 weeks had a score of 14 or less, and subsequently defined a score of less than 15 as a recovery cutoff. In the same study, patients defined as having persistent pain had a score of 28 or greater. Miettinen³⁰ used a recovery cutoff of 20/50 at 3 years.

Floor-Ceiling Effect Floor and ceiling effects have practical clinical relevance, as they represent patients for whom pain and disability estimates may be invalid and for whom changes may not be measurable. Few studies have specifically addressed this issue, and none have specifically analyzed how the MDC varies over the spectrum of possible NDI scores. Riddle et al³⁸ reported that 6% of patients demonstrated a ceiling effect when using the NDI, whereas Hains et al¹⁷ and Vos et al⁴⁹ reported a floor-ceiling effect on the items regarding “personal care” and “concentration.” Some authors report that it is difficult to observe change in scores at the extreme ends of the scale; that is, for both “low” or “normal” scores, as well as for scores that fall in the “severe disability” interval.^{34,38} For example, if a subject with a severe pathology scored 48 out of 50 on the NDI, it may be difficult to objectively distinguish further decline in their status, as they may have attained a near maximum score.

Reliability There is considerable evidence to support appropriate retest reliability of the NDI in both acute and chronic populations, where reliability coefficients above 0.90 have been reported in a number of studies. In contrast, recent large high-quality studies indicated lower reliability^{9,10} (TABLE 2). Possible reasons for variation in reliability between studies are random differences in samples, actual difference in clinical subpopulations, study process, and definitions of “stable.”

TABLE 5

**SUMMARY OF RESPONSIVENESS,
CULTURAL ADAPTATION, RESPONSE BURDEN,
ADMINISTRATIVE (CONTINUED)**

Psychometric properties	Data Extracted
Longitudinal validity correlation with other change scores	<p>Comparisons:</p> <ul style="list-style-type: none"> • GPE, 0.40, 95% CI⁵³ • AUC, 0.79, 95% CI⁵³ • Pearson r, 0.38.⁴⁹ • ROC curve, 0.76.⁴⁹ • Correlation NDI change to NBQ 0.6 and VAS 0.5 <p>AUC, 0.83⁹ ROC, 3.5 (pooled); MID 10.5; sensitivity, 90%; specificity, 70% With COM,³⁶ 0.76 Correlation to Disability Rating Index, 0.95; SF-36 physical, -0.88; VAS activity, 0.86; VAS pain, 0.60¹ Correlation of NDI change to GRC change, 0.3⁴⁴ Correlation to clinicians prediction of change, 0.54⁵⁰ Correlation to PSFS, $r = 0.73$-0.83 change scores⁵⁰</p>
Cultural adaptation	<p>Language/cultural translation/validation Turkish,²¹ Spanish,²² Iranian,³¹ Portugese-Brazilian,¹¹ French⁵³ Agreement between translated version 69% identical answers, ICC = 0.88 Translated into Greek, with input from the developer⁴⁴ Translations available on the MAPI website www.proqoulid.com: English for Australia, English for the United States, English for the UK, Danish, Dutch, Finnish, French, French Canadian, French for Switzerland, German, German for Switzerland, Italian, Italian for Switzerland, Norwegian, Polish, Portuguese, Spanish for Spain, Spanish for the United States Correlation of English and Turkish NDI the r for test and retest was 0.66 and 0.73³</p>
Response burden	<p>Time to complete: 9 min,²¹ 4 min,²² 5 min (stated, not clear how/if measured)²⁹ “fulfilled in 6 min 08 s (54 s) by those with middle-high cultural level, and in 7 min 59 s (1 min 26 s) by those with low one ($p < 0.001$)”²</p>
Administrative burden	5 min to analyze ¹⁸
<p><i>Abbreviations: AUC, operating characteristic curve; CID, clinically important difference; COM, core outcome measure; CWOM, core whiplash outcome measure; ES, effect size; GPE, global perceived effect; GRC, global rating of change; NDI, neck disability index; NPDI, Neck Pain Disability Index; NPQ, Northwick Park Neck Pain Questionnaire; PSFS, Pain-Specific Functional Scale; ROC, receiver operating characteristic; Rx, treatment; SF-36 physical, Short-Form 36 physical component scale; SRM, standardized response mean; VAS, visual analog scale</i></p>	

same study administered 7 different versions of the NDI, and concluded that changing the order of the questions does not change the validity of the tool. Ackelman and Lindgren¹ included individuals with acute and chronic conditions, and demonstrated that there is a higher correlation between the VAS and items on the NDI pertaining to pain and activity in an acute population. Overall, the content of the NDI has demonstrated validity in evaluation of pain and disability in acute and chronic neck conditions, whether from a musculoskeletal or neural source, and whether stemming from a traumatic

or nontraumatic injury.

The NDI was developed as a unidimensional disability instrument for patients with neck pain. Cronbach's α has generally exceeded .85, which is consistent with values reported for other unidimensional scales. However, this statistic alone provides a relatively weak indication of structural validity or dimensionality. Structural validity has also been evaluated using factor analysis in a small number of studies with inconsistent findings. Some studies have supported the NDI as a 1-dimensional scale, while others suggest it has 2 factors. Hains et al¹⁷ observed that all

items are positively correlated with the total score (0.54-0.84). Interitem correlations perform similarly well; the items of “work” and “driving” had the strongest correlation (0.77), and items of “personal care” and “headaches” had the weakest correlation (0.33).¹⁷ Conversely, Wlodyka-Demaille et al⁵³ extracted 2 domains in the French version, “Functional Disability” and “Neck pain,” suggesting that the scale may have 2 dimensions (pain and function) in some subgroups. Similarly, Nieto et al³³ conducted a factor analysis in 150 patients with whiplash injury and found 2 factors that they termed “pain and interference with cognitive functioning” and “functional disability.” This 1- or 2-factor effect has been observed on other brief disability scales which include questions about pain and disability treated as one-dimensional and may reflect the extent to which pain and disability are separate concepts across different pathologies and samples.

A highly powered Rasch analysis of the NDI ($n = 521$ patients) was available for appraisal, based on a manuscript now in press that was shared by the authors.⁴⁵ An advantage of this approach is that it formally assesses whether an instrument satisfies rules for interval scale measurement by examining fit with a Rasch model. In particular, this approach is able to determine whether response options are ordered and might be reasonably interpreted as numbers. This study determined that there was a lack of fit of NDI to the model and that 4 items had disordered response thresholds. Items on headaches and recreation had significant deviation from model expectations. Items on lifting, personal care, and work also demonstrated disordered thresholds. Differential item functioning was detected, indicating that some items did not have the same meaning across subgroups. Unidimensionality tests showed that different subsets of items gave significantly different person estimates, suggesting that the NDI items did not contribute to a single underlying construct. The item on headaches did not fit with the trait

TABLE 6

ENGLISH LANGUAGE (SPECIFIC-TO-NECK) AND 3 (NONSPECIFIC VALIDATED FOR THE NECK) SCALES REVIEWED IN A PREVIOUS SYSTEMATIC REVIEW³⁷

Neck Disability Index 1991
Northwick Park Neck Pain Questionnaire 1994
Disability Rating Index 1994
Patient-Specific Functional Scale 1998
Copenhagen Neck Functional Disability Scale 1998
Neck Pain and Disability Visual Analog Scale 1999
Functional Rating Index 2001
Extended Aberdeen Spine Pain Scale 2001
Bournemouth Neck Questionnaire 2002
Cervical Spine Outcomes Questionnaire 2002
Whiplash Disability Questionnaire 2004

captured by other items. They concluded that the NDI is not unidimensional, but that a revised 8-item NDI would satisfy these requirements and provide interval-level measurement of neck disability. The 8-item NDI was consistent with the 10-item version and had similar correlations to measurements of neck pain intensity.

Responsiveness The NDI appears to have good responsiveness when considering the typical reported effect sizes or standardized response mean (SRM), varying from 0.60 when measured in the total cohort of subjects with “mildly” disabling chronic whiplash-associated disorders, to 0.95 when measured in only the improved cohort (TABLE 5).^{5,23,42} This indicates that the NDI has good ability to detect change over time. Only a few studies calculated the clinically important difference (CID) (TABLE 5).^{10,43,47} When neck pain is of musculoskeletal origin, a CID can be said to have occurred when there is a change of greater than 5 points,⁴³ compared to a 7-point change in patients with neck pain of neural origin.¹⁰

DISCUSSION

THIS STUDY SYNTHESIZED CURRENT research in 37 studies addressing the psychometric properties of the NDI and was able to provide some clinical recommendations regarding its use,

within the limits prescribed by the available evidence. Overall, there is moderate to strong evidence for a spectrum of psychometric properties supporting use of the NDI in patients with acute or chronic neck pain with symptoms of musculoskeletal or neurogenic origin. The relative importance of different psychometric properties will vary according to purpose. For example, when using the NDI to evaluate clinical change in individual patients, the absolute measurement error and responsiveness should be considered most relevant. Conversely, when using the NDI to differentiate different levels of disability, known group validity, a form of discriminative validation that tests differences between known subgroups, would be more important.

There is adequate evidence that the NDI is stable over short test-retest time intervals (0-3 days), although it is less likely that patients defined as stable will have stable scores in longer test-retest intervals. In short-term test-retest intervals researchers often assume that most untreated patients remain relatively stable over 1 to 3 days. In longer test-retest intervals it is more necessary to establish the patient has remained stable. This is often based on a -3 to +3 score on global rating of change instrument. Thus, by definition, patients with small to moderate self-reported changes are defined as

stable in this analysis. Conversely, ICCs reported when the test-retest was based on a single occasion, with the NDI administered in different languages³¹ or different question order, have been exceptionally high (>0.90). This suggests that lower test-retest reliability across longer test intervals may reflect changes in scores due to the episodic nature of neck pain, early recovery, and how “stable” is defined within a given time window. Some studies have used as much as a 5-week interval in patients defined as stable to determine test-retest reliability. While it is important to understand short-term and long-term stability of clinical measurements, it should be appreciated that an increasing number of factors may affect scores as test-retest intervals are extended. Even if a patient’s severity of pain or task difficulty remains consistent over the course of several weeks, the experience of pain and disability will be mediated by psychosocial influences and calibrated against multiple internal and external references. Thus, changes in factors like coping, self-efficacy, or social support may contribute to alterations in perceived disability over a 5-week period. Recalibration of the disability experience in the absence of changes in pain intensity is a plausible explanation for lower estimates of reliability in studies with larger test-retest intervals. However, it is important to consider that these variables do contribute to measurement error in clinical studies. Therefore, reliability studies with longer test-retest intervals may be important to consider when designing clinical research studies and reliability studies, whereas shorter intervals may be appropriate for consideration when estimating error across shorter clinical re-evaluations. The reliability data currently published in the literature provide estimates that reflect different definitions of stable, in different clinical subgroups, with different test-retest intervals allowing users to select estimates most similar to their situation and use.

Only a few studies presented more clinically relevant indicators of measure-

ment error, the SEM or the MDC, both being expressed in unit of the original score and based on the reliability coefficient and the sample variability. The developer of the NDI suggests that MDC is 5 points. The highest estimate comes from Cleland et al,¹⁰ who reported an MDC of 10.2 points for their patient population with cervical radiculopathy, and the lowest was from Voss et al,⁴⁹ who reported an MDC of 1.66 for their study, which included stable patients with recurrent neck pain. Despite a range of 2 to 10, the more common estimate for MDC is around 5/50 (10%). The reasons for this large range of MDC values are largely unclear, although either larger standard deviations or lower reliability coefficients can increase the value of the MDC, so it is not surprising that the larger MDC estimates come from studies with lower ICCs. Across the different studies, MDC variations reflect the effects of test-retest interval and the definition of stability on reliability estimates. Despite variability, the value of 5/50 or 10% commonly used in clinical situations appears to be appropriate for most clinical comparisons that tend to occur over 2 weeks or less.

While the NDI has been shown to be responsive, estimates are highly variable among studies, suggesting that a number of fully powered studies that vary on factors such as the type of intervention, length of follow-up, comorbidity, and the nature of the neck condition will be required to provide more precise estimates in different clinical circumstances. This review did not attempt to evaluate different instruments and, therefore, was not designed to comment on whether the NDI is the most responsive neck scale. However, none of the studies that performed head-to-head comparisons of different neck disability scales indicated that another instrument was superior to the NDI in terms of responsiveness.

There was agreement across studies that the NDI is easy to read and understand, in both its original English format, as well as in subsequent translations. In general, published translations used at

least some of the recommended procedures for valid translation and demonstrated equivalence.⁴ Due to the brief nature of the questionnaire, the NDI has minimal administrative burden, although scoring variations can create potential for confusion. Ceiling/floor effects have been suggested as a potential problem, although there is no consistent definition of what constitutes a ceiling or floor. If one assumes that the MDC is usually at 5 points (and up to 10 points), then scores of 40 or higher and 10 or lower might be considered problematic for detecting worsening or improvement, respectively. Evidence suggests that smaller changes are relevant at these ends of the scale. We would also suggest that within these ranges, clinicians should consider supplementing the NDI with other instruments, like the Patient-Specific Functional Scale,^{10,50} which is able to sample items that are of high or low difficulty, thus, making the scale more amenable to change for patients with this type of clinical presentation.

One limitation of our review stems from the lack of agreed upon quality criteria for synthesis process for psychometric studies. Neither previous systematic review incorporated critical appraisal. Although the first author of this review has addressed the critical appraisal of individual studies by developing a tool for this purpose,^{25,26} there is no clear method to synthesize the extracted psychometric evidence. In some systematic reviews only high-quality studies are synthesized. However, when evaluating an outcome measure, it is important to see how the instrument performs across different contexts and purposes. Furthermore, there are no levels of evidence that create clear categories for study quality. Therefore, we rank ordered studies by quality to allow the reader and ourselves a mechanism to place greater emphasis on the findings from high-quality studies. We summarized the information on psychometrics and usefulness by adapting and expanding a framework used by others. While we tried to make the process

as objective as possible, there are inherently subjective elements, as study results must be taken within the context of the study population, interventions, and purpose, making it difficult to synthesize results from individual studies into global recommendations.

A second limitation in our review is that the scope of our search retrieved full-text papers written in only English or French. We don't expect this limitation to have a substantial impact on our results, as the majority of translation and validation articles were printed in English, and we were able to extract data from English abstracts in non-English text.

Overall, the NDI has a number of features that suggest that it has good clinical utility. These include its brevity and the fact that it has been translated into a number of languages and its responsiveness to detection of clinical change. It is important to have outcome measures that can be applied across different cultural or language subgroups. Although this cross-validation is well under way, future studies may focus on whether the pain/disability experience in neck pain varies across the subgroups and how these variations are reflected on the NDI responses.

It is certainly true that no other instrument has undergone sufficient development or validation to replace the NDI as the instrument of choice in routine evaluation of neck disability. The instrument is suitable for both clinical research and practice, although evidence and measurement principles suggest that patients who score at the extremes may benefit from supplemental scales, like the Patient-Specific Functional Scale. It is an important consideration that introducing new instruments into clinical practice and research requires tremendous efforts in translation/cultural adaptation and knowledge translation. For these reasons, any suggestions of change to different instruments or in the NDI itself should be balanced with this consideration. The most recent suggestion is that an 8-item version of the

NDI performs as well as the original NDI and can be considered as interval data.⁴⁵ An advantage of this suggestion is that it is easier to change to a reduced-item instrument than a different one. Finally, a practical issue for clinicians is the difficulty of obtaining official translations, as these are not usually published with validation studies. Because these are not published with the validation manuscript, it is the responsibility of the clinician to seek out these tools by contacting developers. However, moves to improve the accessibility of translations would benefit clinical practice.

While the NDI may be considered the “gold standard” among outcome measures, this review suggests further investigation is needed. Because the NDI was not developed using a clinimetric process and original pilot testing was conducted on a very small sample, it is not clear whether the NDI captures all of the important concepts for patients with neck pain or weighs pain and disability according to their relative priority. This leaves open the possibility that the addition of items might enhance performance. Others have suggested that removal of items might improve performance.⁴⁵ Currently, variations in scoring exist in the literature; thus, caution should be exercised when presenting or comparing scores as to whether scores are out of 50 points or 100%.

Perhaps, most importantly, there are gaps in defining clinically useful comparative data and benchmarks. For example, the work on MDC and CID is sparse and inconsistent. Defining detectable change and important difference for different clinical subgroups and the impact of different prognostic variables on these would allow clinicians to provide more accurate prognosis and outcome evaluation. These would be important in assisting clinicians in using the NDI to set short- and long-term goals, and to communicate more effectively with payers. Importantly, the current benchmarks set for normality and for levels of disability are arbitrary, and specific data analyses are needed to establish valid benchmarks.

RECOMMENDATIONS FOR FUTURE DEVELOPMENT OF THE NDI

1. Future studies on the psychometric properties in different clinical groups are warranted and should specifically state hypotheses that are to be evaluated, including the specific psychometric property being analyzed, as well as the expected outcome.
2. Studies that determine SEM, MDC, and CID in different subgroups are needed to resolve confusion in the literature and provide more accurate prognostication/outcome evaluation.
3. Authors should consider the availability of their translations and provide mechanisms for these to be obtained by interested users when publishing official manuscripts related to the translation process.
4. Studies that determine NDI benchmarks that clinicians could use with confidence to indicate “normal” versus different levels of disability, and targets for important milestones, like return to work, are needed.
5. Studies that compare the proposed revisions to the 8-item NDI in different clinical contexts and in longitudinal studies are required to determine the need for this revision.
6. Qualitative studies and cognitive interviewing that evaluate how patients respond to items of the NDI would inform our understanding of self-reported disability as reflected on the NDI.

CONCLUSION/KEY POINTS

1. The NDI is reliable, valid, and responsive in numerous patient populations, including patients with acute and chronic conditions, as well as those suffering from neck pain associated with musculoskeletal dysfunction, whiplash-associated disorders, and cervical radiculopathy.^{1,23,43}
2. The NDI should be scored out of 50 as recommended by the developer; a

percentage score used to adjust for unanswered questions must be divided by 2 to restore the expected metric. Caution should be used when reading clinical reports to ascertain what metric was used.

3. Published cultural validation has been performed for the NDI in the following languages: English, French, Iranian, Dutch, Korean, Swedish, and Portuguese.^{1,11,24,49,53} In addition, commercial sites provide translations in English for Australia, English for the United States, English for the UK, Danish, Dutch, Finnish, French, French Canadian, French for Switzerland, German, German for Switzerland, Italian, Italian for Switzerland, Norwegian, Polish, Portuguese, Spanish for Spain, and Spanish for the United States.
4. The MDC is accepted as 5 points, but varies considerably across studies and may be up to 10 for cervical radiculopathy. Short-term therapy goals should require a minimum of 5 points change for whiplash-associated disorder I and II, and up to 10 points for whiplash-associated disorder III; this should be re-evaluated within a timeframe where this change could be anticipated (2 weeks).^{1,10,48,49}
5. The CID is approximately 7 points, so longer-term treatment goals should be set for a minimum 7-point reduction in score to demonstrate treatment benefits.^{10,43}
6. Patients who score in the range of either 40 to 50 or 0 to 10 on the NDI should be considered as approaching a ceiling/floor effect (45/5, respectively), making it difficult to detect subsequent worsening or improvement. When the baseline score is outside of the 10-to-40 range, supplementation of the NDI with a Patient-Specific Functional Scale should be considered.
7. Clinicians may relate an NDI score to the subjective categories reported in the literature, but should include a statement reflecting that these disability benchmarks have not been sufficiently validated nor occupational

demands considered.

A “normal” score of between 0 to 20 points, which represents “none to mild disability,” has been suggested by Miettinen et al.³⁰ Vernon and Mior⁴⁸ suggest that a score between 0 and 4 represents no disability, 5 and 14 mild disability, 15 and 24 moderate disability, 25 and 34 severe disability, and greater than 35 complete disability. Sterling et al⁴¹ suggest that individuals who have recovered have a NDI score of 8 or less, those with mild disability have a score of 10 to 28, and those with moderate to severe disability have a score greater than 30. ●

REFERENCES

1. Ackelman BH, Lindgren U. Validity and reliability of a modified version of the neck disability index. *J Rehabil Med*. 2002;34:284-287.
2. Andrade Ortega JA, Delgado Martinez AD, Al-mecija Ruiz R. [Validation of a Spanish version of the Neck Disability Index]. *Med Clin (Barc)*. 2008;130:85-89.
3. Aslan E, Karaduwan A, Yakut Y, Aras B, Simsek IE, Yagly N. The cultural adaptation, reliability and validity of neck disability index in patients with neck pain: a Turkish version study. *Spine*. 2008;33:E362-365. <http://dx.doi.org/10.1097/BRS.0b013e31817144e1>
4. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*. 2000;25:3186-3191.
5. Bolton JE. Sensitivity and specificity of outcome measures in patients with neck pain: detecting clinically significant improvement. *Spine*. 2004;29:2410-2417; discussion 2418.
6. Bovim G, Schrader H, Sand T. Neck pain in the general population. *Spine*. 1994;19:1307-1309.
7. Chan Ci En M, Clair DA, Edmondston SJ. Validity of the Neck Disability Index and Neck Pain and Disability Scale for measuring disability associated with chronic, non-traumatic neck pain. *Man Ther*. 2008;<http://dx.doi.org/10.1016/j.math.2008.07.005>
8. Chok B, Gomez E. The reliability and application of the Neck Disability Index in physiotherapy. *Physiotherapy Singapore*. 2000;3:16-19.
9. Cleland JA, Childs JD, Whitman JM. Psychometric properties of the Neck Disability Index and Numeric Pain Rating Scale in patients with mechanical neck pain. *Arch Phys Med Rehabil*. 2008;89:69-74. <http://dx.doi.org/10.1016/j.apmr.2007.08.126>
10. Cleland JA, Fritz JM, Whitman JM, Palmer JA. The reliability and construct validity of the Neck Disability Index and patient specific functional scale in patients with cervical radiculopathy. *Spine*. 2006;31:598-602. <http://dx.doi.org/10.1097/01.brs.0000201241.90914.22>
11. Cook C, Richardson JK, Braga L, et al. Cross-cultural adaptation and validation of the Brazilian Portuguese version of the Neck Disability Index and Neck Pain and Disability Scale. *Spine*. 2006;31:1621-1627. <http://dx.doi.org/10.1097/01.brs.0000221989.53069.16>
12. Croft PR, Lewis M, Papageorgiou AC, et al. Risk factors for neck pain: a longitudinal study in the general population. *Pain*. 2001;93:317-325.
13. Eecheute C, Vaes P, Van Aerschot L, Asman S, Duquet W. The clinimetric qualities of patient-assessed instruments for measuring chronic ankle instability: a systematic review. *BMC Musculoskelet Disord*. 2007;8:6. <http://dx.doi.org/10.1186/1471-2474-8-6>
14. Foster GA, Goldsmith CH. Horizon Estimation Using SAS Software. *SAS Global Forum 2008*. San Antonio, TX: 2008.
15. Gay RE, Madson TJ, Cieslak KR. Comparison of the Neck Disability Index and the Neck Bournemouth Questionnaire in a sample of patients with chronic uncomplicated neck pain. *J Manipulative Physiol Ther*. 2007;30:259-262. <http://dx.doi.org/10.1016/j.jmpt.2007.03.009>
16. Goolkasian P, Wheeler AH, Gretz SS. The neck pain and disability scale: test-retest reliability and construct validity. *Clin J Pain*. 2002;18:245-250.
17. Hains F, Waalen J, Mior S. Psychometric properties of the neck disability index. *J Manipulative Physiol Ther*. 1998;21:75-80.
18. Heijmans WPG, Schipholt HJA, Elvers JWH, Oostendorp RAB. Neck disability index Dutch version (NDI-DV): investigation of reliability in patients with chronic whiplash. *Nederlands Tijdschrift Voor Fysiotherapie*. 2002;112:94-99.
19. Hoving JL, O'Leary EF, Niere KR, Green S, Buchbinder R. Validity of the neck disability index, Northwick Park neck pain questionnaire, and problem elicitation technique for measuring disability associated with whiplash-associated disorders. *Pain*. 2003;102:273-281.
20. Humphreys H, Bolton JE. Documenting outcomes in neck pain patients. *J Whiplash Rel Disorders*. 2002;1:5-22.
21. Kose G, Hegguler S, Atamaz F, Oder G. A comparison of four disability scales for Turkish patients with neck pain. *J Rehabil Med*. 2007;39:358-362. <http://dx.doi.org/10.2340/16501977-0060>
22. Kovacs FM, Bago J, Royuela A, et al. Psychometric characteristics of the Spanish version of instruments to measure neck pain disability. *BMC Musculoskelet Disord*. 2008;9:42. <http://dx.doi.org/10.1186/1471-2474-9-42>
23. Kumbhare DA, Balsor B, Parkinson WL, et al. Measurement of cervical flexor endurance following whiplash. *Disabil Rehabil*. 2005;27:801-807. <http://dx.doi.org/10.1080/09638280400020615>
24. Lee H, Nicholson LL, Adams RD, Maher CG, Halaki M, Bae SS. Development and psychometric testing of Korean language versions of 4 neck pain and disability questionnaires. *Spine*. 2006;31:1841-1845. <http://dx.doi.org/10.1097/01.brs.0000227268.35035.a5>
25. MacDermid JC. Critical appraisal of study quality for psychometric articles, evaluation form. In: Law M, MacDermid JC, eds. *Evidence-Based Rehabilitation*. Thorofare, NJ: Slack Inc; 2008:387-388.
26. MacDermid JC. Critical appraisal of study quality for psychometric articles, interpretation guide. In: Law M, MacDermid JC, eds. *Evidence-Based Rehabilitation*. Thorofare, NJ: Slack Inc; 2008:389-392.
27. MacDermid JC, Stratford P. Applying evidence on outcome measures to hand therapy practice. *J Hand Ther*. 2004;17:165-173. <http://dx.doi.org/10.1197/j.jht.2004.02.005>
28. Makela M, Heliövaara M, Sievers K, Impivaara O, Knekt P, Aromaa A. Prevalence, determinants, and consequences of chronic neck pain in Finland. *Am J Epidemiol*. 1991;134:1356-1367.
29. McCarthy MJ, Grevitt MP, Silcocks P, Hobbs G. The reliability of the Vernon and Mior neck disability index, and its validity compared with the short form-36 health survey questionnaire. *Eur Spine J*. 2007;16:2111-2117. <http://dx.doi.org/10.1007/s00586-007-0503-y>
30. Miettinen T, Leino E, Airaksinen O, Lindgren KA. The possibility to use simple validated questionnaires to predict long-term health problems after whiplash injury. *Spine*. 2004;29:E47-51.
31. Mousavi SJ, Parnianpour M, Montazeri A, et al. Translation and validation study of the Iranian versions of the Neck Disability Index and the Neck Pain and Disability Scale. *Spine*. 2007;32:E825-831. <http://dx.doi.org/10.1097/BRS.0b013e31815ce6dd>
32. Nederhand MJ, Ijzerman MJ, Hermens HJ, Turk DC, Zilvold G. Predictive value of fear avoidance in developing chronic neck pain disability: consequences for clinical decision making. *Arch Phys Med Rehabil*. 2004;85:496-501.
33. Nieto R, Miro J, Huguet A. Disability in subacute whiplash patients: usefulness of the neck disability index. *Spine*. 2008;33:E630-635. <http://dx.doi.org/10.1097/BRS.0b013e31817eb836>
34. Pietrobon R, Coeytaux RR, Carey TS, Richardson WJ, DeVellis RF. Standard scales for measurement of functional outcome for cervical pain or dysfunction: a systematic review. *Spine*. 2002;27:515-522.
35. Pool JJ, Ostelo RW, Hoving JL, Bouter LM, de Vet HC. Minimal clinically important change of the Neck Disability Index and the Numerical Rating Scale for patients with neck pain. *Spine*. 2007;32:3047-3051. <http://dx.doi.org/10.1097/BRS.0b013e31815cf75b>
36. Rebbeck TJ, Refshauge KM, Maher CG, Stewart M. Evaluation of the core outcome measure in whiplash. *Spine*. 2007;32:696-702. <http://dx.doi.org/10.1097/01.brs.0000257595.75367.52>
37. Resnick DN. Subjective outcome assessments for cervical spine pathology: a narrative review. *J Chiro Med*. 2005;4:113-134.
38. Riddle DL, Stratford PW. Use of generic versus region-specific functional status measures on

[LITERATURE REVIEW]

patients with cervical spine disorders. *Phys Ther*. 1998;78:951-963.

39. Skolasky RL, Riley LH, 3rd, Albert TJ. Psychometric properties of the Cervical Spine Outcomes Questionnaire and its relationship to standard assessment tools used in spine research. *Spine J*. 2007;7:174-179. <http://dx.doi.org/10.1016/j.spinee.2006.07.005>
40. Sterling M, Jull G, Kenardy J. Physical and psychological factors maintain long-term predictive capacity post-whiplash injury. *Pain*. 2006;122:102-108. <http://dx.doi.org/10.1016/j.pain.2006.01.014>
41. Sterling M, Jull G, Vicenzino B, Kenardy J. Characterization of acute whiplash-associated disorders. *Spine*. 2004;29:182-188. <http://dx.doi.org/10.1097/01.BRS.0000105535.12598.AE>
42. Stewart M, Maher CG, Refshauge KM, Bogduk N, Nicholas M. Responsiveness of pain and disability measures for chronic whiplash. *Spine*. 2007;32:580-585. <http://dx.doi.org/10.1097/01.brs.0000256380.71056.6d>
43. Stratford PW, Riddle DL, Binkley JM, Spadoni G, Westaway MD, Padfield B. Using the Neck Disability Index to make decisions concerning individual patients. *Physiother Canada*. 1999;51:107-112.

44. Trouli MN, Vernon HT, Kakavelakis KN, Antonopoulou MD, Paganas AN, Lionis CD. Translation of the Neck Disability Index and validation of the Greek version in a sample of neck pain patients. *BMC Musculoskelet Disord*. 2008;9:106. <http://dx.doi.org/10.1186/1471-2474-9-106>
45. van der Velde G, Beaton D, Hogg-Johnston S, Hurwitz E, Tennant A. Rasch analysis provides new insights into the measurement properties of the neck disability index. *Arthritis Rheum*. 2009;61:544-551. <http://dx.doi.org/10.1002/art.24399>
46. Vernon H. The Neck Disability Index: state-of-the-art, 1991-2008. *J Manipulative Physiol Ther*. 2008;31:491-502. <http://dx.doi.org/10.1016/j.jmpt.2008.08.006>
47. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. *J Manipulative Physiol Ther*. 1991;14:409-415.
48. Vernon HT. Assessment of self-rated disability, impairment and sincerity of effort in whiplash-associated disorder. *J Musckel Pain*. 2000;8:155-167.
49. Vos CJ, Verhagen AP, Koes BW. Reliability and responsiveness of the Dutch version of the Neck Disability Index in patients with acute neck pain in general practice. *Eur Spine J*. 2006;15:1729-

1736. <http://dx.doi.org/10.1007/s00586-006-0119-7>

50. Westaway MD, Stratford PW, Binkley JM. The patient-specific functional scale: validation of its use in persons with neck dysfunction. *J Orthop Sports Phys Ther*. 1998;27:331-338.
51. White P, Lewth G, Prescott P. The core outcomes for neck pain: validation of a new outcome measure. *Spine*. 2004;29:1923-1930.
52. Wlodyka-Demaille S, Poiraudreau S, Catanzariti JF, Rannou F, Fermanian J, Revel M. The ability to change of three questionnaires for neck pain. *Joint Bone Spine*. 2004;71:317-326. <http://dx.doi.org/10.1016/j.jbspin.2003.04.004>
53. Wlodyka-Demaille S, Poiraudreau S, Catanzariti JF, Rannou F, Fermanian J, Revel M. French translation and validation of 3 functional disability scales for neck pain. *Arch Phys Med Rehabil*. 2002;83:376-382.



MORE INFORMATION
WWW.JOSPT.ORG

EARN CEUs With JOSPT's Read for Credit Program

JOSPT's **Read for Credit (RFC)** program invites *Journal* readers to study and analyze selected *JOSPT* articles and successfully complete online quizzes about them for continuing education credit. To participate in the program:

1. Go to www.jospt.org and click the link in the “**Read for Credit**” box in the right-hand column of the home page.
2. Choose an article to study and when ready, click “**Take Exam**” for that article.
3. Login and pay for the quiz by credit card.
4. Take the quiz.
5. Evaluate the RFC experience and receive a personalized certificate of continuing education credits.

The RFC program offers you 2 opportunities to pass the quiz. You may review all of your answers—including the questions you missed. You receive **0.2 CEUs** for each quiz passed, and the *Journal* website maintains a history of the quizzes you have taken and the credits and certificates you have been awarded in the “**My CEUs**” section of your “**My JOSPT**” account.

APPENDIX A

NECK DISABILITY INDEX

This questionnaire is designed to help us better understand how your neck pain affects your ability to manage everyday-life activities. Please mark in each section the one box that applies to you, although you may consider that two of the statements in any one section relate to you. Please mark the box that **most closely** describes your present-day situation.

Section 1: Pain Intensity

- I have no neck pain at the moment.
- The pain is very mild at the moment.
- The pain is moderate at the moment.
- The pain is fairly severe at the moment.
- The pain is very severe at the moment.
- The pain is the worst imaginable at the moment.

Section 2: Personal Care

- I can look after myself normally without causing extra neck pain.
- I can look after myself normally, but it causes extra neck pain.
- It is painful to look after myself, and I am slow and careful.
- I need some help but manage most of my personal care.
- I need help every day in most aspects of self-care.
- I do not get dressed. I wash with difficulty and stay in bed.

Section 3: Lifting

- I can lift heavy weights without causing extra neck pain.
- I can lift heavy weights, but it gives me extra neck pain.
- Neck pain prevents me from lifting heavy weights off the floor but I can manage if items are conveniently positioned, ie. on a table.
- Neck pain prevents me from lifting heavy weights, but I can manage light weights if they are conveniently positioned.
- I can lift only very light weights.
- I cannot lift or carry anything at all.

Section 4: Work

- I can do as much work as I want.
- I can only do my usual work, but no more.
- I can do most of my usual work, but no more.
- I can't do my usual work.
- I can hardly do any work at all.
- I can't do any work at all.

Section 5: Headaches

- I have no headaches at all.
- I have slight headaches that come infrequently.
- I have moderate headaches that come infrequently.
- I have moderate headaches that come frequently.
- I have severe headaches that come frequently.
- I have headaches almost all the time.

Section 6: Concentration

- I can concentrate fully without difficulty.
- I can concentrate fully with slight difficulty.
- I have a fair degree of difficulty concentrating.
- I have a lot of difficulty concentrating.
- I have a great deal of difficulty concentrating.
- I can't concentrate at all.

Section 7: Sleeping

- I have no trouble sleeping.
- My sleep is slightly disturbed for less than 1 hour.
- My sleep is mildly disturbed for up to 1-2 hours.
- My sleep is moderately disturbed for up to 2-3 hours.
- My sleep is greatly disturbed for up to 3-5 hours.
- My sleep is completely disturbed for up to 5-7 hours.

Section 8: Driving

- I can drive my car without neck pain.
- I can drive my car with only slight neck pain.
- I can drive as long as I want with moderate neck pain.
- I can't drive as long as I want because of moderate neck pain.
- I can hardly drive at all because of severe neck pain.
- I can't drive my car at all because of neck pain.

Section 9: Reading

- I can read as much as I want with no neck pain.
- I can read as much as I want with slight neck pain.
- I can read as much as I want with moderate neck pain.
- I can't read as much as I want because of moderate neck pain.
- I can't read as much as I want because of severe neck pain.
- I can't read at all.

Section 10: Recreation

- I have no neck pain during all recreational activities.
- I have some neck pain with all recreational activities.
- I have some neck pain with a few recreational activities.
- I have neck pain with most recreational activities.
- I can hardly do recreational activities due to neck pain.
- I can't do any recreational activities due to neck pain.

Patient Name _____ Date _____

Score _____ [50]

Copyright: Vernon H & Hagino C, 1991. For permission to use the NDI, please contact Dr. Howard Vernon at hvernon@cmcc.ca

[LITERATURE REVIEW]

APPENDIX B

DATA EXTRACTION FORM FOR STUDIES EVALUATING THE PSYCHOMETRICS OF OUTCOME MEASURES

Authors: _____ Year: _____ Rater: _____

Instructions

When using the data extraction form, it is important to realize that the purpose of data extraction is to remove or extract the specific information reported by authors within a study, not to evaluate the validity or value of that piece of information. To make data extraction as useful as possible and to avoid the need for repeated data extractions, read the accompanying guide and be as specific as possible when extracting information.

Data Extracted	
Population studied	
Population	
Intervention	
Reliability	
Reliability (relative)	
Reliability (absolute)	
Minimum detectable change (MDC)	
Content/structural validity	
Internal consistency	
Content validity	
Floor/ceiling effects	
Factorial validity	
Item response/Rasch analyses	

Construct/criterion validity	
Known groups	
Convergent	
Divergent	
Longitudinal validity	
Concurrent criterion	
Predictive criterion	
Responsiveness/clinical change	
Responsiveness	
Minimally clinical important difference (MCID)	
Usefulness/practicality	
Readability	
Interpretability	
Time to administer	
Administration burden	
Cultural applicability	

© JC MacDermid 2008

[LITERATURE REVIEW]

DATA EXTRACTION GUIDE FOR STUDIES EVALUATING THE PSYCHOMETRICS OF OUTCOME MEASURES

Instructions

Psychometric studies may evaluate a wide spectrum of potential psychometric properties and aspects that relate to the utility or usability of outcome measures. Studies will not address every aspect. In summarizing what is known about the psychometrics and utility of outcome measures, it is useful to collect information within standardized domains so that this information might be more easily synthesized to provide a summary of our collective knowledge about psychometrics and the utility of any given outcome measure. To this end, the following guide has been devised to describe the separate properties that might be evaluated in a psychometric study. Listed here are an explanation of each property and how it might be measured within a study. The accompanying extraction form can then be used to collect specific information about these psychometric or utility properties from particular studies.

When using the data extraction form, it is important to realize that the purpose of data extraction is to remove or extract the specific information reported by authors within a study, not to evaluate the validity or value of that information. Evaluation of the quality of articles (also called critical appraisal) is performed in a separate step. To make data extraction as useful as possible and to avoid the need for repeated data extractions, it is best to be as specific as possible when extracting information.

There is no clear or easy method to synthesize this information. Based on the findings of extraction, you may elect to present the synthesized data in a descriptive way by creating a summary table in each category. If you find some studies with similar designs, you may be able to conduct a meta-analysis of some properties such as minimal detectable change (MDC), clinically important difference (CID), or standard error of measurement (SEM).

	Property	Methods That Might Be Used to Collect This Information/Relevant Statistics
Population studied		
Population	A description of the study population	Report meaningful demographics and indicators of the population studied: sample size, pathology/disorder, demographics, setting, acute versus chronic, how and where subjects were chosen.
Intervention	Interventions (if applicable) applied for longitudinal studies	Description of the nature, frequency, and intensity of the intervention, and the follow-up interval
Reliability		
Reliability description	The extent to which the same results are obtained on repeated administrations of the same measure when no change in status has occurred (reliability) or how precise the scores are on repeated measurements (agreement). Relative reliability evaluates the extent to which variations on repeated assessment of the same individual compare to the variability between individuals.	Test procedures or measures are typically reapplied on repeated occasions in individuals considered to have a stable condition during that timeframe in which repeated testing occurs. Repeated testing may be performed on different occasions (test-retest) for self-report measures, OR by the same rater (intrarater), or different raters (interrater) if a server-based scale is used. In some cases, different test instruments (inter-instrument) are evaluated. The most common statistic used is the intraclass correlation coefficient (ICC) for quantitative data and Kappa for nominal data. SEM is used to present a quantitative estimate of the reliability in the original units of measure. Some consider that sufficient time between tests is within 48 hours for acute conditions, and 4–14 days for chronic conditions. Report the type of reliability evaluated and coefficients obtained.
Reliability (relative)	The extent to which variability in test scores on repeated tests of the same person are related to variability overall (including between people)	ICCs and their associated confidence intervals

Reliability (absolute)	The extent to which the same results are obtained on repeated administrations of the same measure when no change in status has occurred (reliability) or how precise the scores are on repeated measurements (agreement). Specific quantitative estimates of reliability in individual units portray the actual amount of variation expected between repeated applications—reported in the original units of measure.	This may be reported as <ol style="list-style-type: none"> SEM (possibly expressed as coefficient of variation in older articles) Altman and Bland graphical technique where the difference on repeated tests for each individual is plotted versus their mean score (and the overall and limits of 2 standard deviations are shown)
MDC	Based on the reliability and defined level of confidence, this measure reflects the amount required to change before achieving a level of confidence that exceeds the random error that occurs in stable patients.	Extract the number and level of confidence
Content/structural validity		
Internal consistency	The extent to which items on a test or subscale are related (an indication of the consistency of the concept measured)	Cronbach's alpha is the inter-item correlation usually reported. Report alpha and whether it relates to the entire instrument or specific subscales.
Content validity	The extent to which the domain of interest is adequately sampled by the items in the measure. In assessing content validity, it is important to consider the population to which the measure applies, the completeness of the content, the content's relevance, and emphasis.	A variety of techniques can be used to assess the extent to which items on a given measure reflect the necessary content to capture the concept of interest. Some of the techniques typically found are listed. Extract what was done and what was found. <ol style="list-style-type: none"> Patients and experts were involved during item selection/reduction. Patients were consulted for reading and comprehension. Cognitive interviews were done with patients to determine the meaning of items. Expert panels or Delphi procedures were used to select items or evaluate the validity of the instrument. During translation, specific study of the meaning of the questions to another cultural or language group was studied.
Floor-ceiling effects	When the measure fails to demonstrate a worse score in patients who clinically deteriorated and/or an improved score in patients who clinically improved	Descriptive statistics of the distribution of scores were presented graphically or numerically and indicate this. Report the percentage of patients who sustained a floor or ceiling effect. Note that different studies may define floor/ceiling differently, so extract how these attributes were determined as well as the percentage.
Factorial validity	The extent to which factor analysis supports assumptions surrounding constructs measured as defined by the measure or as indicated by subscale structure	Factor analysis is conducted and compared to the inherent structure of the instrument or factor analysis on which its construction is based. Report the number of factors derived and the extent to which these findings agree with the instrument structure or original factor structure.

[LITERATURE REVIEW]

Item response/ Rasch analyses	The extent to which items cross a range of difficulty, or a spectrum of the concept measured	Items can be placed along a spectrum using item response theory or Rasch analysis. Analyses might address item difficulty, individual ability curves, and comparison of ability estimation. Most commonly, the order of item placement in a test and/or the discriminative ability of specific items are defined. Report whether these analyses were conducted and if they determined that items crossed a range of difficulty. Rasch analysis can be used to evaluate differential item functioning (DIF)—report if done and whether items demonstrated DIF.
Construct/criterion validity		
Construct validity description	Constructs are artificial frameworks that are not directly observable. Construct validity is the extent to which measures perform according to a priori defined constructs. Construct validity can be cross-sectional or longitudinal (predictive). Constructed hypotheses can assess convergent validity where measures are thought to represent similar constructs or divergent validity where it is assumed they measure different constructs.	The extent to which scores relate to other measures in a manner consistent with theoretically derived hypothesis concerning the domains that are measured OR the expected differences between “known groups” OR in relation to other indices that are similar (convergent) or different (divergent). In each case, hypotheses are formulated. Results are evaluated to determine whether they are acceptable in accordance with the hypotheses.
Known groups	Known groups are constructed on theoretical assumptions that they should be different, and the difference is assessed.	Extract the groups, their mean scores, and level of significance
Convergent	Relationship between similar scales/tests	Extract test names and correlations; when summarizing, consider grouping into strong (>0.70) and moderate (0.40–0.70) correlations
Divergent	Relationship to scales/tests assumed unrelated	Extract test names and correlations
Longitudinal validity	Relationship between change scores among similar scales/tests measured on different time points where change is expected	Extract test names and correlations
Criterion validity description	Criterion validation is determined by comparing a given outcome measure to an accepted standard of measure. For subjective constructs such as pain and disability, it is sometimes argued that there is no criterion and, therefore, validation focuses on construct validity. In other cases, the term comparability is used when scales measure the same subjective construct, particularly if one is considered more accepted or rigorous. Concurrent validity is assessed by comparing a scale and its criterion at a single point in time. Predictive validity is evaluated by determining the extent to which the results of administering an outcome measure at one point in time is associated with a future status or outcome.	Authors will state that their measure is being compared against a specific instrument and report the correlation or agreement between the measures. Extract the test names and correlations.
Concurrent criterion	Comparing a scale and its criterion at a single point in time assesses concurrent validity.	Extract the test names and correlations
Predictive criterion	Predictive validity is evaluated by determining the extent to which the results of administering an outcome measure at 1 point in time is associated with a future status or outcome.	Extract the test names and correlations and time interval

Responsiveness/clinical change		
Responsiveness	The ability to detect important change over time in the concept being measured. Subjects are evaluated at 2 points in time and change in patients who had improved (or have improved an important amount) is compared across different measures (and/or to patients who have not improved). Hypotheses are formulated and results are in agreement.	Extract indicators of responsiveness, including effect size, standard response mean, and the method for assessing whether patients were improved, stable, or worse.
CID	The difference in scores in the domain of interests that patients perceive to be beneficial mandates a change in patients' management. Information is provided about what difference in score would be clinically meaningful. How clinically meaningful is defined may vary (global rating of change with different methods of rating or establishing cutoffs for importance are often used).	Extract the minimally important difference (MID), minimally clinical important difference (MCID), or CID and the method/cut-off used to define importance. MID, MCID, and CID are 3 different terms for the same concept.
Usefulness/practicality		
Readability	The questionnaire is understandable by all patients. Authors provide specific information on readability as evaluated by the target population or specific tools are applied to evaluate readability (eg, grade level can be assigned in some software packages)	State method and results obtained
Interpretability	The degree to which one can assign qualitative meaning to quantitative scores or define subgroups by scores. Authors provide information on the interpretation of scores: 1. Comparative data for relevant subgroups (acute versus chronic, etc) 2. Validation of subjective categories of outcome: excellent/good/fair; normal/abnormal	State the type of categorization and relevant scores and, if appropriate, the method used to validate/analyze subgroup differences
Time to administer	Authors provide specific information on time to administer.	Extract time
Administration burden	Ease of method used to calculate the questionnaire's score. Authors provide specific information on the ease of scoring. If this is not specifically reported in the text of the study, it cannot be extracted. However, during a systematic review, compare scoring algorithms and administrative burden and report this. This must be done in a systematic way and the method should be documented. For example, to extract the specific scoring algorithm for an instrument and compare it to others, multiple raters should evaluate the relative complexity of these scoring algorithms.	Extract study data on scoring method burden

[LITERATURE REVIEW]

Cultural applicability	<p>The extent to which an instrument contains items that will be meaningful across different cultural groups or subgroups for which the measure is relevant or may be applied. Cross-cultural adaptation may be used to translate measures into different languages and to convert any items having specific meaning to culturally relevant items.</p> <p>In cross-cultural adaptation, a number of analyses may be performed, including a structured translation process that includes forward and backward translation and retesting reliability of the translated version, and a formal evaluation of the meaning of items in different cultures. Highlight the results of cross-cultural adaptation as well as any languages/cultural adaptations that have been reported. If specific reliability and validity values are reported, they can be documented in the appropriate data collection boxes above; however, an annotation should be made to note that the psychometric property applies to an adapted version.</p>	Extract language/cultural translation performed (and relevant findings)
------------------------	---	---

© JC MacDermid 2008

CRITICAL APPRAISAL OF STUDY DESIGN FOR PSYCHOMETRIC ARTICLES: EVALUATION FORM

Authors: _____ Year: _____ Rater: _____

Evaluation Criteria	Score		
	2	1	0
Study question			
1. Was the relevant background research cited to define what is currently known about the psychometric properties of the measures under study and the need or potential contributions of the current research question?			
Study design			
2. Were appropriate inclusion/exclusion criteria defined?			
3. Were specific psychometric hypotheses identified?			
4. Was an appropriate scope of psychometric properties considered?			
5. Was an appropriate sample size used?			
6. Was appropriate retention/follow-up obtained? (Studies involving retesting or follow-up only)			
Measurements			
7. Documentation: Were specific descriptions provided or referenced that explain the measures and their correct application/interpretation (to a standard that would allow replication)?			
8. Standardized methods: Were administration and application of measurement techniques within the study standardized, and did they consider potential sources of error/misinterpretation?			
Analyses			
9. Were analyses conducted for each specific hypothesis or purpose?			
10. Were appropriate statistical tests conducted to obtain point estimates of the psychometric property?			
11. Were appropriate ancillary analyses done to describe properties beyond the point estimates (confidence intervals [CI], benchmark comparisons, standard error of measurement [SEM], minimally important difference [MID])?			

[LITERATURE REVIEW]

Recommendations			
12. Were the conclusions/clinical recommendations supported by the study objectives, analysis, and results?			
Subtotals (of columns 1 and 2)			
Total score % (sum of subtotals/24 × 100), or, if for a specific paper an item is deemed inappropriate, then sum items, divide by 2 times the number of items, and multiply by 100 to get the percentage score.			

© JC MacDermid 2008

CRITICAL APPRAISAL OF STUDY DESIGN FOR PSYCHOMETRIC ARTICLES

Interpretation Guide

To decide which score to provide for each item on a quality checklist, read the following descriptors. Pick the descriptor that sounds most like the study being evaluated with respect to a given item. If there is no documentation of an action, treat it as not done.

Question	Score	Descriptors
Study question		
1	2	The authors: 1. Performed a thorough literature review, indicating what is currently known about the psychometric properties of the instruments or tests under study from previous research studies. 2. Presented a critical and unbiased view of the current state of knowledge. 3. Indicated how the current research question evolves from a gap in the current knowledge base. 4. Established a research question based on the above.
	1	All of the above criteria were not fulfilled (little reference to previous research and present gaps in knowledge), but a clear rationale was provided for the research question.
	0	A foundation for the current research question was not clear, and the rationale was not founded on previous literature.
Study design		
2	2	Specific inclusion/exclusion criteria for the study were defined, the practice setting was described, and appropriate demographic information was presented, yielding a study group generalizable to a clinical situation.
	1	Some, but not all information on participants and place is provided. For example, age/sex/diagnosis and the name or type of the practice is given, but without additional information. Information on the type of patients is briefly defined, but is insufficient to allow the reader to generalize the study to a specific population.
	0	No information on the type of clinical settings or study participants is provided.
3	2	Authors identified specific hypotheses, which included the specific type of reliability (intra/interrater or test-retest) or validity (construct/criterion/content, longitudinal/concurrent, convergent/divergent) being tested. A priori hypothesis definitions are provided of level of reliability and for validity, as well as expected relationships (strength of associations) or constructs.
	1	Types of reliability and validity being tested were stated, but not clearly defined in terms of specific hypotheses.
	0	Specific types of reliability or validity under evaluation were not clearly defined, and specific hypotheses on reliability and validity were not stated. ("The purpose of this study was to investigate the reliability and validity of..." can be rated as zero if no further detail on the types of reliability and validity or the nature of specific hypotheses is provided.)
4	2	An appropriate scope of psychometric properties would be indicated by: 1. A detailed focus on reliability that includes multiple forms of reliability (at least 2 of intrarater, interrater, test-retest); as well as both relative and absolute reliability [eg, intraclass correlation coefficients (ICC), standard error of measurement (SEM), and minimally important difference (MID)] 2. A detailed focus on validity that includes multiple forms of validity (content) judgmental (structured, eg, expert review/survey or qualitative interviews) or statistical (eg, factor analyses), construct (known group differences; convergent/divergent associations), criterion (concurrent/predictive), responsiveness, and established predictive, evaluative or discriminative properties 3. Some aspects of both reliability and validity were examined concurrently.
	1	Two or more psychometric properties were evaluated, however, scope was narrow and did not meet the above criteria (eg, internal consistency and 1 test/form of validity).
	0	The scope of psychometric properties was narrow as indicated by evaluation of only 1 form of reliability or validity.

[LITERATURE REVIEW]

5	2	Authors performed a sample size calculation and obtained their recruitment targets. Post-hoc power analyses and/or confidence intervals (CI) confirm that the sample size was sufficient to define relatively precise estimates of reliability or validity.
	1	The authors provided a rationale for the number of subjects included in the study, but did not present specific sample size calculations or post-hoc power analyses. For simple reliability/validity statistics if sample is greater than 100 but without justification for sample size.
	0	Size of the sample was not rationalized or is clearly underpowered.
6	2	Ninety percent or more of the patients enrolled for study were reevaluated.
	1	More than 70% of the patients eligible for study were reevaluated.
	0	Less than 70% of the patients eligible for study were reevaluated.
Measurements		
7	2	The authors provided or referenced a published manual/article that outlines specific procedures for administration, scoring (including how scoring algorithms handle missing data), and interpretation of test procedures that includes any necessary information about positioning/active participation of the subject, any special equipment required, calibration of equipment if necessary, training required, cost, and examiner procedures/actions. If no manual is referenced, then the text describes the procedures in sufficient detail so they can be replicated.
	1	Procedures are referenced without any details, or a limited description of procedures is included in the text.
	0	Minimal description of procedures is provided and without appropriate references.
8	2	All of the measurements, including test administration and scoring, were performed in a standardized way. For self-report, this is characterized by a statement of who administered the forms and by what process (definition of rules for exclusion of forms [eg, too many missing items] or individuals [language, comprehension, etc]). For impairment measures, this includes calibration of any equipment, use of consistent measurement tools and scoring, a priori exclusion of any participants likely to give invalid results or unable to complete testing (not exclusion after enrollment), use of standardized instructions, and test procedures.
	1	No obvious sources of bias in how tests were performed/administered, but minimal attention or description of the extent to which the above standards were maintained.
	0	No description of the extent to which the above standards were maintained or an obvious source of bias in data collection methods.
Analyses		
9	2	Authors clearly defined which specific analyses were conducted for each of the stated specific hypotheses of the study. This may be accomplished through organization of the results under specific subheadings, or by demarcating which analyses addressed specific psychometric properties. Data were presented for each hypothesis/research question.
	1	Data were presented for each hypothesis, but authors did not clearly link analyses to hypotheses.
	0	Data were not presented for each hypothesis or psychometric property outlined in the purposes or methods.

10	2	Appropriate statistical tests were conducted: 1. Reliability (Relative; ICCs for quantitative data, Kappa for nominal data); absolute ; SEM or plot of score differences versus average score showing mean and 2 standard deviation (SD) limit—Altman and Bland) 2. Clinical relevance—minimal detectable change (MDC), MID 3. Validity a. Validity associations—Pearson correlations for normally distributed data, Spearman rank correlations for ordinal data, or other correlations if appropriate b. Validity tests of significant difference—an appropriate global test such as analysis of variance was used where indicated, with post-hoc tests that adjusted for multiple testing 4. Responsiveness—standardized response means or effect sizes or other recognized responsiveness indices were used
	1	Appropriate statistical tests were used in some instances, but suboptimal choices were made in other analyses.
	0	Inappropriate use of statistical tests
11	2	For key indicators such as reliability coefficients indices, at least 2 of the following were presented: 1. Appropriate confidence intervals (CI) 2. Comparison to appropriate benchmarks or standards 3. SEM Correlation matrices for validity analysis may not require that each individual correlation be presented with its associated CI; however, CIs and benchmarks should be used according to standards for this type of analysis.
	1	Either CIs or appropriate benchmarks were used, but not both.
	0	Benchmarks or CIs were either used inappropriately or not included.
Recommendations		
12	2	Authors made specific conclusions and clinical recommendations that were clearly related to the specific hypotheses stated at the beginning of the study and supported by the data presented.
	1	Authors made conclusions and clinical recommendations that were general, but basically supported by the study data, OR authors made conclusions and clinical recommendations for only some of the study hypotheses.
	0	Authors made vague conclusions without any clinical recommendations, and conclusions OR recommendations contradicted the actual data presented.

© JC MacDermid 2008