Kristian Pylkäs

PSYCHOMETRIC PROPERTIES OF NECK DISABILITY INDEX - A SYSTEMATIC REVIEW AND META-ANALYSIS

Syventävien opintojen kirjallinen työ

Syyslukukausi 2024

Kristian Pylkäs

PSYCHOMETRIC PROPERTIES OF NECK DISABILITY INDEX - A SYSTEMATIC REVIEW AND META-ANALYSIS

Kliininen laitos

Syyslukukausi 2024

Vastuuhenkilöt: Mikhail Saltychev, Juhani Juhola

TURUN YLIOPISTO
Lääketieteellinen tiedekunta

PYLKÄS, KRISTIAN: Psychometric properties of neck disability index - a systematic review and meta-analysis

Syventävien opintojen kirjallinen työ, 32 sivua
Fysiatria
Lokakuu 2024

The Neck Disability Index (NDI) is a self-report questionnaire that measures disability associated with the activities of daily living due to neck pain. It is widely used by both researchers and clinicians. Despite the popularity and wide use of the NDI, the comprehensive analyses of its psychometric properties have either been lacking or are outdated. The aim of this study was to evaluate the data available on the psychometric properties of the NDI, particularly from a clinical perspective.

This study was a systematic review and meta-analysis based on a literature search of the Medline, Embase, PsychINFO, Web Of Science and Scopus. The search resulted in 492 records. After the screening, 79 of these records were included into analysis. 70 studies were considered to be of low risk of systematic bias. Alpha was >0.81. Pooled test-retest intraclass correlation coefficient was 0.91. The NDI correlations with pain rating scacles varied from 0.38 to 0.89. 13 studies found the NDI to be unidimensional and 15 - two- or three dimensional. The minimal detectable change varied from 3 % to 27 % and minimal clinically important difference from 5 % to 33 %. Pooled area under the curve was 0.74. Most studies have not detected floor or ceiling effect. Sex-related differential item functioning has been present in one study.

The NDI demonstrated good internal consistency and test-retest reliability without floor or ceiling effect. In most situations, the NDI could be considered a unidimensional scale. The NDI well correlated with the common scales of pain and disability. The minimal clinically important difference and minimal detectable change were around 15 % (7.5/50 points).

Keywords: Neck Pain, Psychometrics, Disability Evaluation, Reproducibility of Results

This work is based on the original publication: Mikhail Saltychev, Kristian Pylkäs, Aleksandra Karklins & Juhani Juhola (19 Jan 2024): Psychometric properties of neck disability index – a systematic review and meta-analysis, Disability and Rehabilitation, DOI: 10.1080/09638288.2024.2304644

Lääketieteellinen tiedekunta

PYLKÄS, KRISTIAN: Niskakipuindeksin psykometriset ominaisuudet - systemaattinen katsaus ja meta-analyysi

Syventävien opintojen kirjallinen työ, 32 sivua
Fysiatria
Lokakuu 2024

Niskakipuindeksi (NDI) on kysely, joka mittaa niskakivun aiheuttamaa haittaa päivittäistoimintoihin. Se on laajalti käytössä sekä tutkijoiden että kliinikkojen keskuudessa. NDI:n suosiosta ja laajasta käytöstä huolimatta, kattavat analyysit sen psykometrisista ominaisuuksista ovat joko puuttuneet tai ovat vanhentuneita. Tämän tutkimuksen tavoitteena oli arvioida saatavilla olevaa tietoa NDI:n psykometrisista ominaisuuksista, erityisesti kliinisestä näkökulmasta.

Tämä tutkimus oli systemaattinen katsaus ja meta-analyysi, joka perustui Medlinen, Embasen, PsychINFOn, Web Of Sciencen ja Scopusin kirjallisuushakuun. Haku tuotti 492 tulosta. Seulonnan jälkeen näistä 79 sisällytettiin analyysiin. 70 tutkimusta katsottiin olevan vähäisen systemaattisen harhan riskin alaisia.

NDI:n eri osa-alueiden sisäinen yhteneväisyys oli hyvä, alfa >0.81. Yhteenlaskettu toistettavuuden luokansisäisen korrelaation kerroin (ICC) oli 0.91. NDI:n korrelaatio kipujanan kanssa vaihteli välillä 0.38–0.89. 13:ssa tutkimuksessa NDI:n todettiin olevan yksiulotteinen ja 15:ssa tutkimuksessa kaksi- tai kolmiulotteinen. Minimimuutos, jonka NDI pystyy havaitsemaan (MDC), vaihteli välillä 3 %–27 %, ja pienin kliinisesti merkitsevä ero välillä 5 %–33 % (keskimäärin 15 %). ROC-käyrän alla olevan alueen pinta-ala oli 0,74. Suurin osa tutkimuksista ei ole havainnut lattia- tai kattovaikutusta. Yhdessä tutkimuksessa on havaittu NDI:n ominaisuuksien vaihtelevan riippuen vastaajan sukupuolesta.

Avainsanat: Niskakipu, Psykometriikka, Toimintakyvyn arviointi, Tulosten toistettavuus

Tämä työ perustuu alkuperäiseen julkaisuun: Mikhail Saltychev, Kristian Pylkäs, Aleksandra Karklins & Juhani Juhola (19 Jan 2024): Psychometric properties of neck disability index – a systematic review and meta-analysis, Disability and Rehabilitation, DOI: 10.1080/09638288.2024.2304644

Table of contents

1 INTRODUCTION

In 1980, Dr. J.C. Fairbank introduced the Oswestry Low Back Pain Disability Questionnaire, today known as the Oswestry Disability Index (ODI) [1]. The ODI has soon became a popular scale to evaluate the severity of disability caused by low back pain. In 1991, Dr. H. Vernon modified the ODI in order to use it among people with neck pain [2]. Since then, the Neck Disability Index (NDI) has become widespread among both researchers and clinicians [3]. Several translations of the NDI have been introduced [3]. A rough search at Medline, based only on titles and abstracts, showed that while in 1991-2000 ten studies have employed the NDI, in 2011-2020 there have been 1,870 such publications. This trend seems to continue, as there have already been over 800 publications mentioning the NDI in the abstracts since 2021 until this day. This is not a surprise, as the NDI is short, easy and free to use, and its properties have seemed to be good across very different settings [4]. During these three decades, several modifications of the NDI have been suggested, though failing to replace the original version [3, 4].

Despite the popularity of the NDI, there have been only a few attempts to summarize the available evidence on its psychometric properties. In 2008, Dr. H. Vernon himself has conducted a narrative review of the NDI psychometrics [3]. The records had been retrieved using Science Citation Index from 1991 to 2007 and 22 publications have been analysed. That review has stated that the psychometric properties of the NDI are well established in numerous cultural groups and the NDI is highly reliable, strongly internally consistent and unidimensional scale with excellent convergent and divergent validity. Also, a minimal clinically important difference (MCID) has been set at 3/50 to 5/50 points. In 2009, MacDermid et al. have conducted the first, and probably ever since the last, systematic review on the psychometrics of the NDI [5]. They have reviewed 37 papers grading most of the studies as having low risk of systematic bias. It has been noticed that the included reports had usually approached the problem from a theoretical point of view, serving scientific research and focusing on such properties of the NDI as construct validity or group reliability. Instead, the clinical relevancy of the NDI had less been studied. The intraclass correlation coefficients (ICCs) have ranged from 0.50 to 0.98 and most of the studies had reported the NDI unidimensionality. The minimum detectable change (MDC) has been set at 5/50 points, while the MCID has been set between 5/50 and 19/50 points. The convergent validity has been good. Another review has come close to the present topic [6]. In 2015, Yao et al. have investigated the cross-cultural adaptations of the NDI in order to provide some advice to improve the further translations. At that point, 14 different translations have been identified. The internal consistency of the NDI has been found acceptable, and the ICC has been >=0.70. Only one of the included studies had reported a floor effect. That review has also drew attention to the fact that clinical interpretability of the NDI had been left outside the scope of most of the research. Other earlier reviews on the topic have mostly been narrative, focusing primarily on other issues regarding the NDI than psychometrics. In 2011, Howell et al. have conducted a

narrative review on the role of the NDI in evaluating association between neck pain and cervical range of motion [7]. The popularity of the NDI among surgeons operating on cervical spine could be seen as several narrative reviews have been conducted to summarise the evidence on the use of the NDI among patients undergoing cervical surgery [8-10].

This way, it seems that despite the wide use of the NDI for both clinical and research purposes, the comprehensive analyses of its psychometric properties have been missing or outdated. The objective of the present review was to evaluate the data available on the psychometric properties of the NDI, especially from the clinical point of view.

2 METHODS

2.1 Inclusion and exclusion criteria (PICO)

A study was considered relevant if the studied population were adults (>=17 years) with neck pain including radiating pain in upper extremities. A study was considered relevant if the NDI has been used as a main or secondary outcome measure. Reports of any design and settings published in peer- reviewed academic journals with abstracts available were included. Conference proceedings, unpublished data, case reports, case series (n<=10), theses and respective were excluded. A study was excluded if the health conditions of studied population involved 'red flags' – specific causes of neck pain like cancer, infection, acute trauma, osteoporotic cervical fracture, congenital neck deformities etc. A study was included if it had reported an outcome of interest – any numerical data concerning the following psychometric properties of NDI: content and face validity; reliability (internal consistency [alpha], test-retest correlation, inter-rater reliability (intraclass correlation coefficient [ICC]); construct validity (exploratory [EFA] and confirmatory [CFA] factor analyses); criterion validity including convergent, concurrent and predictive validity; Item Response Theory (IRT)- or Rasch-based estimates; or clinical relevancy (minimal clinically important difference [MCID], minimal detectable change [MDC], or floor- and ceiling effects).

2.2 Data sources and search strategy

Medline, Embase, PsychINFO, Web of Science, and Scopus were searched in April 2023. The search was conducted by one of the researchers (MS). The search strategy is presented in Supplement 1. The references of identified articles and reviews were also checked for relevancy.

2.3 Selection strategy

The records identified from the data sources were stored in the Endnote software (Endnote X7.8, Thomson Reuters) (MS). Using a build-in search engine of the Endnote software, duplicates, conference proceedings, theses, reviews, case reports etc. were deleted (MS). Two independent reviewer teams were formed: MS vs. KP and JJ vs. AK. These reviewer teams screened the titles and abstracts of the remaining articles and assessed the full texts of potentially relevant papers (figure 1). Disagreements between the reviewers were planned to resolve by a consensus or by a third reviewer, (no actual need for a third referee has appeared during the process).

2.4 Extraction strategy

The data needed for a quantitative assessment were extracted using a standardized form based on recommendations by the Cochrane Handbook for Systematic Reviews of Interventions. The formncluded: the name of first author, the year of publication, country, language, group sizes, sex and age distributions, main complaints, and the estimates of main outcomes.

2.5 Assessment of risk of systematic bias

The Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies by National Heart, Lung, and Blood Institute (https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools) was modified to correspond the needs of this particular review with its specific goals. Of 14 items included in the original tool, seven items are related to exposure (items #6-#10, "12 and #14) and one to the loss to follow-up (item #13). As the exposure to an intervention was irrelevant in this review and the majority of included studies was expected to be of a cross-sectional design, the following six criteria were included:

1. Was the research question or objective in this paper clearly stated?
2. Was the study population clearly specified and defined?
3. Was the participation rate of eligible persons at least 50%?
4. Were all the subjects selected or recruited from the same or similar populations (including the same time period)? Were inclusion and exclusion criteria for being in the study prespecified and applied uniformly to all participants?
5. Was a sample size justification, power description, or variance and effect estimates provided?
6. Were the outcome measures (dependent variables) clearly defined, valid, reliable, and implemented consistently across all study participants?

Each out of six criterion was graded by an independent reviewer as "low", "high" or "unclear" risk of systematic bias. The reviewers were also asked to judge the total risk as "low" or "high".

2.6 Statistical analysis (meta-analysis)

When only range for the NDI score had been reported, a mean was approximated as 'mean ≈ (max + min)/2'. When variance had been reported as a range or as an interquartile range (IQR), a standard deviation (SD) was calculated as: 'SD ≈ IQR/1.35' or 'SD ≈ range/4'. The strength of correlation was interpreted as follows: 0.90 to 1.00 'very high'; 0.70 to 0.90 'high'; 0.50 to 0.70 'moderate'; 0.30 to 0.50 'low'; and 0.00 to 0.30 'negligible' [11]. The Cohen's kappa was interpreted as follows: values ≤ 0 'no agreement'; 0.01–0.20 'none to slight'; 0.21–0.40 'fair'; 0.41– 0.60 'moderate'; 0.61–0.80 'substantial'; and 0.81–1.00 'almost perfect agreement' [12]. The intraclass correlation coefficient (ICC) describes the strength of test-retest reliability. In this review, the following cut-offs for the ICC were used: <0.5 'poor'; 0.5 – 0.75 'moderate'; 0.75 – 0.9 'good'; and > 0.90 'excellent reliability' [13]. The Cronbach's alpha was interpreted as: >=0.90 'excellent'; 0.80 – 0.89 'good'; 0.70 – 0.79 'acceptable'; 0.60 – 0.69 'questionable'; 0.50 – 0.59 'poor'; and <0.5 'unacceptable' [14].

In this review, the area under the curve (AUC) indicated how well the NDI was able to distinguish those who have improved over time from those who have not. The AUC estimates were pooled to summarize the overall diagnostic accuracy of the test. The AUC takes values from 0.0 'perfectly inaccurate' to 1.0 'perfectly accurate

test'. In this study, the AUC of 0.7 to 0.8 was considered 'acceptable', 0.8 to 0.9 'excellent', and > 0.9 was considered 'outstanding' [15].

A random-effects model was used assuming the presence of substantial heterogeneity across the identified studies. Pooled estimates were accompanied by 95% confidence intervals (95% CIs) and two- tailed p-value, when appropriate. When conducting the analysis of publication bias, a one-tailed p- value was reported. The level of significance was set at <0.05. The test for heterogeneity was conducted using the Q test considering heterogeneity being present if Q was greater than the degree of freedom (number of studies – 1). The I^2 statistic described the percentage of the variability in effect estimates that was due to heterogeneity rather than sampling error (chance). A potential publication bias was evaluated graphically by using a funnel plot and by the Egger's test for asymmetry of the funnel plot (test for the Y intercept = 0 from the linear regression of normalized effect estimate against precision). The trim-and-fill method was used to impute studies into funnel plot to correct asymmetry if needed. A publication bias was assessed when the number of studies in the model was >=10. The sensitivity test included removing one study at a time assessing the consequence of that removal on a pooled estimate. Along with 95% CI, prediction intervals (PIs) were calculated as: 95% PI = pooled estimate + 1.96 x Tau. While confidence interval defines the average effect expected to be seen, prediction interval defines the boundaries of true effect expected to be seen for a single new observation (e.g., next patient). All calculations were performed using the Comprehensive Meta- Analysis, Version 4, Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H., Biostat®, Englewood, NJ 2022.

3 RESULTS

3.1 Search results

The search resulted in 492 records (figure 1). After excluding duplicates and irrelevant/ineligible records by using automation tools (Endnote®), 132 records were screened based on their titles and abstracts. At this point, MS and KP achieved a substantial agreement of 94% (Cohen's kappa 0.68), while JJ and AK achieved a fair agreement of 71.2% (Cohen's kappa 0.22). The composite kappa was calculated as an average of the estimates obtained by two reviewer groups resulting in overall moderate kappa of 0.45. Of the selected records, 106 were screened based on their full-texts and 27 records were excluded. The remaining 79 records were reviewed qualitatively including the assessment of risk of systematic bias. The quantitative estimates were retrieved from 77 records, each record representing a different study. However, pooling the estimates was possible for overall 42 studies. The meta-analysis was conducted when numerical data for mean estimates and their variances were available. This way, pooled test-retest intraclass correlation and Area Under the Curve were estimated. All the other variables of interest were subjected to qualitative analysis.

3.2 Descriptive characteristics of selected records

The studies have been published between 2003 and 2023 (table 1). Of them, 18 (23%) have been conducted in the USA. The studies have used 34 different translations, including the original English language questionnaire (n=24, 30%). For many records, the exact sample sizes were hardly definable due to the inconsistency of reports, subgrouping, drop-outs etc. The sample sizes have ranged from 17 to 2,070 respondents. In some cases, when more than one group was reported, a bigger group was selected for further analysis. Some studies have reported sex and age distributions by several subgroups. If so, then the reported group estimates were pooled. Sex distribution has been reported by 76 studies. Of these, 58 (76%) have been predominated by women. The average age has widely varied across the studies from 26 up to 65 years.

While the inclusion criteria have varied, most of the studies (n=51, 61%) have been conducted among patients with some neck pain lasted over three months and without red flags. The second most common studied subpopulation (n=15, 19%) consisted of people, who have undergone cervical surgery or who have been waiting for surgery. Radiculopathy in upper extremities has been an inclusion criterion in four studies [16-19]. Whiplash has been the main complaint in six studies, though a study by Gabel et al. has also included patients with nonspecific neck pain [20-25]. One study has been completed among patients with acute neck pain [26]. One study has included healthy volunteers along with people with neck pain of any duration [27]. One study has included patients with non- defined neck pathologies [28]. A single study has been conducted among healthy volunteers [29]. While that study did not strictly match the inclusion criteria, it was included into a qualitative analysis as it might provide a wider inside concerning the validity of the NDI in diverse situations.

3.3 Assessment of risk of systematic bias

Nine studies were considered to be of high risk of systematic bias and the rest 70 studies were considered to be of low risk of systematic bias (table 2).

Psychometric characteristics reported by the included studies are presented in table 3.

3.4 Content and face validity

As part of translation and adaptation process, 25 studies have evaluated content validity (expert opinion) [18, 27, 28, 30-50] and 26 face validity (patient opinion) [18, 27, 32-34, 36-39, 41-54]. All of these studies have found both the content and face validity of the NDI to be good.

3.5 Reliability (internal consistency)

The Cronbach's alpha, as a measure of internal consistency, has been reported by 34 (43%) studies. Of them, 33 studies have observed good or excellent alpha between 0.81 and 0.97. Only one study has estimated alpha below 0.80 [36]. Only one study has accompanied an alpha with a confidence limit [55].

3.6 Reliability (intraclass correlation coefficient [ICC])

The ICC estimates have been reported by 38 studies (48%). Most of them have accompanied estimates with 95% confidence intervals, which gave an opportunity to perform a meta-analysis on this estimate (figure 2). None of the studies have clarified which one out of 10 possible ICC types has been used. When not reported, the variance (95% CI) has been imputed from a study closest by its mean estimate. Due to the asymmetricity of 95% CIs, they were converted into SDs. The analysis was based on 38 studies. The effect size index was the mean. The random-effects model was employed for the analysis. The mean pooled ICC was excellent 0.91 (95% CI 0.90 to 0.93). The Z-value was 117.0 with $p<0.001$. The Q-value was 730.4 with 37 degrees of freedom and $p<0.001$ signalling that the true effect differed across these studies. The I^2 was 95%, which showed that some 95% of the variance in observed effects reflected variance in true effects rather than sampling error. Tau, the standard deviation of true effect sizes, was 0.04 in raw units. Assuming that the true effects were normally distributed (in raw units), the prediction interval was 0.83 to 0.99 – the true effect size in 95% of all comparable populations falls in this intervalWhile there were signs of publication bias (Egger's regression intercept $p<0.0001$), there was no need for a trim-and-fill correction (figure 3). Removing one study at a time from the model did not significantly affect the pooled estimate.

3.7 Reliability (test-retest correlation coefficient)

Different correlation coefficients (Spearman/Pearson) have been used to describe test-retest reliability by five studies [16, 20, 43, 46, 56]. As only one of them [20] have reported the 95% confidence intervals, a meta-

synthesis on this estimate was not possible. The correlation coefficients were significant and positive varying from 'moderate' 0.56 to 'very high' 0.98.

3.8 Convergent validity

The convergent validity of the NDI has been tested against numerous very different scales. Half of all the articles has reported on correlation between the NDI and either pain numeric rating scale (NRS) or pain visual analogue scale (VAS). The estimates varied from low positive 0.38 to high positive 0.89 (16 reports exceeded 0.50 – the cut-off of moderate correlation).

The Neck Pain and Disability scale (NPAD) has been compared to the NDI by eight studies with positive significant correlations between moderate 0.66 ad high 0.86 [34, 41, 42, 50, 56-59]. Only one study has reported the 95 CI for that correlation [58].

Eight studies have compared the NDI with different scales to assess the level of global improvement (Global Rating of Change, Global Perceived Effect. Patient Global Improvement Change Scale, Global Rating Scale, Patient Global Impression of Change etc.) [33, 36, 38, 59-63]. The correlation coefficients varied from low positive 0.32 to moderate positive 0.60.

Six studies have reported the correlations between the NDI and the Patient-Reported Outcomes Measurement Information System (PROMIS) [64-69]. The reported correlations between the NDI and the PROMIS subscales were subscale varied from low 0.47 for 'Anxiety' and 0.46 for 'Depression' to moderate to high for other subscales: 'Physical function' 0.60. PROMIS Pain Interference (PI): 0.71.0.55 to 0.83, 'Pain interference' 0.71 to 0.74, 'Pain behavior' 0.58, 'Fatigue' 0.56, 'Sleep disturbance' 0.58, 'Social roles' -0.64, and 'Pain intensity' 0.67.Five studies have compared the NDI with the SF-36 [41, 48, 52, 53, 70]. The approaches for that have been different – some studies have reported only one composite estimate while the others have reported correlations for two or even for eight subscales. The SF-36 consists of eight subscales each producing an individual score: physical functioning (PF), role physical (RP), bodily pain (BP), general health (GH), vitality (VT), social functioning (SF), role emotional (RE), and mental health (MH). One study has reported a low correlation of 0.44 between the NDI and the MH and moderate 0.56 with the PF [70]. These correlations have been reported as positive, which has assumingly been a typo as such a correlation could be either negative or, at least, close to zero. Two studies have reported low negative correlations with a single composite of the SF-36 of -0.43 to -0.44 [48, 52]. One study has reported mostly low negative correlations for the seven out of eight subscales of the SF-36: PF -0.36, PR -0.48, BP -0.44, GH -0.25, VT -0.41, SA -0.42, MH -0.37 [53]. One study has reported negligible negative correlations for all eight subscales of the SF-36: PF -0.23, PR -0.19, BP -0.29, GH -0.04, VT 0.02, SA - 0.10, ER -0.16, MH 0.07 [41].

The Patient Health Questionnaire (PHQ-9) has been compared with the NDI by three reports showing positive

correlations from low 0.42 to moderate 0.69 [17, 20, 34].

Two studies have compared the NDI with the Problem Elicitation Technique reporting moderate positive correlation of 0.57 to 0.62 [23, 57]. Two studies have compared the NDI with the Northwick Park Neck Pain Questionnaire reporting positive correlation of moderate 0.56 to high 0.89 [23, 71]. Two studies have compared the NDI with the Tampa Scale for Kinesiophobia reporting low positive correlation from 0.38 to 0.53 (95% CI 0.23 to 0.74) [25, 72].

Several less common scales have been used as comparators by some individual studies: Neck Bournemouth Questionnaire: 0.80 [73], Generalized Anxiety Disorder Scale-7 0.52 [34], Veterans RAND 12-item Physical Component Survey (r= -0.66) [74], Copenhagen Neck Functional Disability Scale (r= 0.87) [40], Hospital Anxiety and Depression Scale (r= 0.42), SF-6D (r= 0.82) [41], Patient-Specific Functional Scale (r= 0.42) [47], Pain Catastrophizing Scale (r= 0.64 [95% CI 0.38 to 0.81]) [72], taking pain medications (r= -0.52) [17], absence from work due to pain (r= 0.64) [17], and frequency of pain (r= 0.29) [17].

A few studies have employed such objective comparators as T1 vertebra slope minus cervical lordosis (r= 0.33) [75], cervical lordosis (r= -0.27) [75], and myelopathy seen on MRI image (r= 0.44) [40].

3.9 Construct validity (Exploratory Factor Analysis [EFA])

Of the studies, 28 have included the EFA. To be precise, this number included studies by Lue et al. and by Santiago-Reynoso et al., which have employed the CFA instead of the EFA in order to determine the number of factors [39, 45]. Of them, 13 have found the NDI to be unidimensional [21, 22, 28, 32, 34, 45, 49, 51, 52, 54, 55, 76, 77], 14 – two-dimensional [17, 24, 30, 31, 33, 36, 38, 39, 41, 44, 46-48, 53] and one study has reported on a three-dimensional structure [37]. Among two-factor models, there were not clear patterns in item distribution – item sets for each of two factors were very different in different studies. Of the studies observing a two-factor structure, four studies have not reported the eigenvalues [17, 36, 39, 41]. The rest 10 studies have demonstrated the same pattern – there has been a very conservative approach to interpret the Kaiser rule for retaining (eigenvalue >1.0) [24, 30, 31, 33, 38, 44, 46-48, 53]. The same trend was seen for all 10 studies – the eigenvalue of the first factor was between 4.0 and 6.0, while the second factor showed an eigenvalue, which was only slightly higher than 1.0, varying between 1.0 and 1.34 (below 1.2 in eight studies).

3.10 Construct validity (Confirmatory Factor Analysis [CFA])

As Lue et al. [39], Santiago-Reynoso et al. [45] and Gabel et al. [22, 78] have employed the CFA in a way that is usually associated with the EFA (defining the number of retaining factors) and Lauridsen et al. [51] have analysed the modified 8-item version of the NDI, only one true CFA on the original NDI was available [55]. In

that study [55], the highest loadings have been seen for items 'pain intensity', 'reading' and 'driving', and the lowest for items 'lifting', 'headaches' and 'sleeping'.

3.11 Clinical relevancy (Minimal Clinically Important Difference [MCID] and Minimal Detectable Change [MDC])

Every third study (n=24) has calculated a minimal detectable change (MDC). Some difficulties were emerged when extracting these estimates, as well as the estimates for a minimal clinically important difference (MCID). Firstly, different statistical formulas have been used. Secondly, the level of statistical significance varied from 0.90 up to 0.99. Thirdly, some studies did not clearly specify if the NDI score was reported as points (0-50) or as percentage (0%-100%). With those reservations, the reported MDC varied widely from 3% to 27% and MCID from 5% to 33%.

3.12 Clinical relevancy (Area Under the Curve [AUC])

Of the studies, 12 have reported AUC estimates. Of them, 10 studies have provided 95% CIs [16, 18, 19, 60, 61, 79-83]. The AUC varied from 0.57 to 0.96. The meta-analysis is based on 11 studies (figure 3). When not reported, the variance (95% CI) has been imputed from a study closest by its mean estimate. One study [59] was excluded from the meta-analysis as there was no variance reported and the mean AUC estimate was far higher than estimate reported by any other study. The effect size index was the mean. The random-effects model was employed for the analysis. The mean pooled AUC was acceptable 0.74 (95% CI of 0.68 to 0.80) – to be precise, the lower confidence limit 0.68 fell just below the limit of an acceptable level of 0.70. The Z-value was 24.2 with $p < 0.001$. The Q-value was 44.1 with 10 degrees of freedom and $p < 0.001$ signalling that the true effect differed across these studies. The I^2 was 77%, which showed that some 77% of the variance in observed effects reflected variance in true effects rather than sampling error. Tau, the standard deviation of true effect sizes, was 0.09 in raw units. Assuming that the true effects were normally distributed (in raw units), the prediction interval was 0.53 to 0.95 – the true effect size in 95% of all comparable populations falls in this interval. While there were signs of publication bias (Egger's regression intercept p=0.006), there was no need for a trim-and-fill correction (figure 4). Removing one study at a time from the model did not significantly affect the pooled estimate.

3.13 Clinical relevancy (floor and ceiling effects)

At least 26 (33%) studies have investigated floor and ceiling effects. Only one study has reported both floor and ceiling effects [29]. Concerning that study, it remained unclear how ceiling effect might be seen in a population consisted of healthy volunteers. Other 25 studies saw neither floor nor ceiling effects of the NDI.

3.14 Item Response Theory and Rasch analyses

Ailliet et al. have reported that the NDI composite score does not show the sex-related DIF, but the DIF has been seen for individual items ‘headaches’, ‘sleep’, ‘lifting’ [30]. Another study has seen the sex- related DIF for items 2, 8 and 9 [84]. The same study has reported that difficulty parameter for all the items was good and discrimination ability ranged from 0.71 to 2.49 [84]. Van Der Velde et al. have employed the Rasch analysis stating that the properties of the item ‘headaches’ differed significantly from the other items [85].

4 DISCUSSION

This systematic review identified 79 observational studies reporting the psychometric properties of the NDI. A meta-analysis was possible for two variables – the ICC and the AUC. The internal consistency of the NDI was good with an alpha between 0.81 and 0.97. Test-retest reliability was overall good showing moderate to very high correlations and an excellent pooled ICC of 0.91. While some studies have reported the two- or even three-dimensional structure of NDI, in most of situations the NDI can be considered a unidimensional scale, which supports the comparability of the NDI scores. The NDI correlated well with such scales as the NPAD and PROMIS, while correlations with pain severity, global improvement, SF-36 and PHQ-9 might significantly vary in different populations. Strong conclusions concerning the convergent validity of the NDI with other scales could not be achieved as the have only been single studies on each scale. It is safe to assume that a change of 15% (7.5/50 points) in the NDI score is the minimal change noticed by a respondent as meaningful (MCID). The NDI was well able to discriminate people who have improved from those who have not. No floor or ceiling effect was detected by 25 studies. There have been a few reports on a potential sex-related DIF for the several NDI items – the NDI score may slightly depend on the sex of respondent. The overall risk of systematic bias was low for most of the studies. Clinical relevancy of the results of the review is presented in brief in table 4.

4.1 Strengths and weaknesses

When there are tens of studies on the subject, even a comprehensive search might, and probably has, let some relevant publications go unrecognized. The identified studies have often described the topic, studied populations, settings, samples and outcome measures in many dissimilar terms, which diminish the preciseness of the findings. In case of multi-group studies, the reviewers were sometimes forced to make voluntaristic decisions on including or excluding some groups into analysis. Variance, needed for the quantitative pooling of the estimates, has often been unreported. In some cases, the variance estimates were calculated using the available data or imputed from other similar studies. The heterogeneity level (also heterogeneity in true effect) was high, widening the confidence and prediction intervals. It should to be taken into account that meta-analysis is always an approximation and should not be directly compared to findings seen in original research. The samples were more or less predominated by women and by middle aged people – it is uncertain if the NDI behaves similarly among younger people or among people older than 60 years. Taking into account all these concerns, a big number of studied identified through a systematic and comprehensive search suggests that the results of this review are reliable enough to identify several trends existing in the body of the NDI literature.

4.2 Findings regarding the NDI internal consistency in relation to previous knowledge

Based on 33 reports, the alpha varied between 0.81 and 0.97. The estimates were comparable with those observed in 2008 and 2009 [3, 5]. Then, based on seven reports, two systematic reviews have seen a high

alpha between 0.74 to 0.93 and 0.70 to 0.96, respectively. In the present review, only one study has reported a 95% confidence limit for an alpha. Without a reported variance, a meta-synthesis was not possible to conduct. Several studies have reported other alpha-related estimates, like changes in alpha when excluding one item at a time resulting in remarkable stability of the original 10-item NDI.

4.3 Findings regarding the NDI repeatability in relation to previous knowledge

The included studies have used two statistical approaches for the task – different correlation coefficients like Spearman, Pearson and the ICC. This is not directly comparable with previous reviews as they have reported r-coefficients and the ICC together. The correlations observed by previous reviews have overall been high (>0.90). In this review, the correlation coefficients varied from moderate 0.56 to very high 0.98. An estimate of 0.56 has been reported by Cleland et al. [16] probably as a result of failed methodology, as appointed by Dr. H. Vernon in his letter in 2008 [4]. Thus, it is safe to assume that the variation seen in the present study was good, being between 0.86 and 0.98. The variability might be caused by very different time intervals between repeated measures employed by the included studies, which might affect the results significantly. E.g., Cook et al. have seen a correlation of 0.97 after one day but only 0.48 after one week [28].

The ICC has been used in half of the included studies. The pooled estimate was excellent 0.91 (95% CI 0.90 to 0.93). Unfortunately, none of the studies has specified what exact type of ICC (out of 10 available) has been used. The common choice for the test-retest ICC is a two-way mixed-effects model [86]. Thus, it is safe to assume that this model has been used by most of the selected studies. However, some uncertainty remained.

4.4 Findings regarding the construct validity of NDI (Exploratory Factor Analysis)

When the most common use of a scale is through its composite score, the unidimensionality (a scale measures no more than a single underlying construct) is an extremely important issue. Surprisingly many of the included publications have reported 'good factor structure' even when unidimensionality was not achieved. When a scale is multidimensional, the composite scores, obtained from two different persons, samples or populations, are not directly comparable as different factors may contribute to a total score differently depending on the particular characteristics of respondents. Of the included studies, 13 have reported the NDI to be unidimensional, 14 – two-dimensional, and one study has observed a three-dimensional structure. When analysing 10 out of 14 reports on a two-dimensional structure, it could be noticed that the difference between the first and the second factors were substantial – eigenvalues over 4.0 versus eigenvalues just above 1.0. Taking into account 13 reports on unidimensionality, the authors feel that the Kaiser's cut-off of 1.0 could be a little relax here. If so, then the NDI could be considered to be a unidimensional scale in most of the situations. To put it simple, eigenvalues represent the importance of individual items regarding a particular factor. The bigger eigenvalue the more important it is to the factor in question. There is no strong consensus on the cut-off for eigenvalues, even though 1.0 is often used as a limit

(Kaiser's rule). The process of retaining factors during a factor analysis is not purely mathematical but rather logical operation, which takes into account the context of analysis and the differences between eigenvalue [87]. E.g., when the eigenvalue of one factor is much larger than the other and this smaller eigenvalue deviates from 1.0 only a little, the researcher has to decide (based on background information) if the difference between two eigenvalues is that large that the second factor could be omitted as unimportant even if its eigenvalue slightly exceeds the usual cut-off.

4.5 Findings regarding the construct validity of NDI (Confirmatory Factor Analysis)

Only one study employed the CFA on the non-modified NDI and for the purpose, which is usually associated with the CFA approach [55]. Based on the distributions of the item loadings, items 'pain intensity', 'reading' and 'driving' were the most important causes of disability, while 'lifting', 'headaches' and 'sleeping' were the least important items. This observation was not comparable with any of the previous studies and, therefore, the need for additional confirmation is clear. It could be speculated that maintaining a slightly flexed head position during reading or driving may exacerbate neck symptoms.

4.6 Findings regarding the clinical relevancy of NDI (MDC and MCID)

The reported MDC and MCID varied widely from 3% to 27% and from 5% to 33%, respectively. Of the included studies, 13 have suggested the MDC of less than 11% (5.5/50 points) and 13 over 11% (5.5/50 points). The variance has not been reported. Here, a weighted mean was calculated, based on each study sample size, resulting in 15% (7.5/50 points). With some reservations, it could be assumed that this level can be used as a cut-off for both the MCID and the MDC. In his review, Dr. H. Vernon has established the MDC at 4% (2/50 points) and the MCID at 7% to 10% (3.5/50 to 5/50 points), even if there had been a report on the MDC over 10% [88]. A review by MacDermid et al. has suggested the MDC of 10% (5/50 points) [5]. The present MDC was substantially higher than one suggested by previous reviews. The reason for that may be a considerably larger set of analysed studies or the use of weighted reported means. The MDC of 15% was very similar to that calculated for the Oswestry Disability Index (the NDI is a modification of the Oswestry Disability Index) – around 14% to 15%.

4.7 Findings regarding the clinical relevancy of NDI (AUC)

The AUC describes the probability of achieving a 'right' result. In other words, the AUC shows how precisely the NDI is able to identify people with improved or worsened levels of disability. The mean pooled AUC was acceptable 0.74 (95% CI of 0.68 to 0.80) – to be precise, the lower confidence limit
0.68 fell just below the limit of an acceptable level of 0.70. A previous review by MacDermid et al. has reported substantially lower confidence limit of 0.50 for the ICC range [5]. This difference may be explained by the larger number of studies identified in the present review. While the previous review has reported the ICC

estimates from two studies, the present review used estimates extracted from 12 studies.

4.8 Findings regarding the clinical relevancy of NDI (floor and ceiling effects)

It is safe to assume that the NDI shows neither floor no ceiling effect in the studied populations. In other word, in most of the cases, the NDI is able to distinguish people with different level of disability through the entire spectrum of scale including those with very severe or only mild disability.

4.9 Findings regarding the Item Response Theory and Rasch analyses of NDI

Only three studies have employed either Item response theory or Rasch analysis. The main finding was that the NDI may behave slightly differently when applied to different sexes. One study has conducted the Rasch analysis [85]. Here, the results were considered uncertain as the main purpose of that study had been to develop a modified version of the NDI.

4.10 Findings regarding the convergent validity of NDI

Numerous scales have been used as comparators to the NDI. Consistently substantial correlations were found between the NDI score and the Neck Pain and Disability scale (NPAD), Patient-Reported Outcomes Measurement Information System (PROMIS), Patient Health Questionnaire (PHQ-9), Problem Elicitation Technique, and Northwick Park Neck Pain Questionnaire. Other scales have been used only once or the results have been inconsistent.

4.11 Suggestions for further research

The basic psychometric research of the NDI is probably going to be expanded by new translations as they appear. Instead, more profound studies on the topic are scarce and there is a clear need for them. Item response theory analyses across different populations and settings are warranted. While the body of literature on the NDI is extensive, research on the convergent validity of the NDI is insufficient and should be continued. So far, there have been only a few publications on the convergent validity of NDI with modern scales based on the International Classification of Functioning, Disability and Health (ICF), like e.g., the WHODAS 2.0. Altogether, assessing the properties of the NDI in the light of the ICF could be a topic, which is very important both for clinicians and researchers. E.g., hardly any study on the NDI has attempted to focus on a functional profile instead of a composite score. While composite score certainly plays an important role for statistical and administrative purposes, one numerical score can hardly describe the entire diversity of functioning on a personal level. Thus, further research may focus on the properties of the individual NDI items instead of being concentrated exclusively on the NDI total score.

4.12 Conclusions

Based on the 79 observational studies of mostly low risk of systematic bias, the NDI demonstrated good internal consistency with an alpha between 0.81 and 0.97, overall good test-retest reliability and no floor or ceiling effect. In most of the situations the NDI could be considered a unidimensional scale. The NDI correlated well with such scales as pain NRS or VAS, NPAD and PROMIS, but, otherwise, the knowledge on the convergent validity was either scarce or inconsistent. Both the MCID and the MDC for the NDI were estimated to be 15% (7.5/50 points). The NDI was well able to discriminate people who have improved from those who have not. The NDI score may slightly depend on the sex of respondent.

5 TABLES AND FIGURES

Table 1. Descriptive characteristics of identified studies.

Study	Country	Language	Participants	N	Women %	Age, years
Ailliet 2013 [30]	Netherland	Dutch	Neck pain	338	66	41
Ailliet 2015 [89]	Netherland	Dutch	Neck pain	337	66	41
Aljinovic 2022 [20]	Croatia	Croatian	Whiplash	30	57	38
Aslan 2008 [56]	Turkey	Turkish	Neck pain	88	74	38
Bakhtadze 2021 [90]	Russia	Russian	Neck pain	136	76	42
Bakhtadze 2015 [31]	Russia	Russian	Neck pain	123	81	38
Bjorklund 2017 [79]	Sweden	Swedish	Neck pain	223	100	53
Carreon 2010 [91]	USA	English	Cervical surgery	682	66	53
Chien 2015 [80]	Taiwan	Chinese	Cervical surgery	45	49	56
Cleland 2008 [16]	USA	English	Neck pain	89	69	42
Cleland 2006 [81]	USA	English	Radiculopathy	17	47	51
Cook 2006 [28]	Brazil	Brazilian Portuguese	Mixed	203	38	43
Cramer 2014 [76]	Germany	German	Neck pain	558	76	50
Croft 2016 [21]	USA	English	Whiplash	123	55	43
Cruz 2015 [32]	Portugal	Portuguese	Neck pain	113	78	52
En 2009 [92]	Australia	English	Neck pain	20	65	65
Farooq 2017 [33]	Pakistan	Urdu	Neck pain	76	61	43
Gabel 2016 [22]	Australia	English	Whiplash/Neck pain	1767	69	40
Gay 2007 [73]	USA	English	Neck pain	23	70	50
Geete 2023[34]	India	Hindi	Neck pain	100	63	36
Geoghegan 2023 [70]	USA	English	Cervical surgery	290	40	49
Goh 2020 [8]	Australia		Cervical surgery	539	46	54
Hoving 2003 [23]	Australia	English	Whiplash	71	83	40
Hung 2018 [93]	USA	English	Neck pain	1945	49	58
Hung 2019 [64]	USA	English	Neck pain	763	50	58
Irmak 2019 [29]	Turkey	Turkish	Healthy	49	41	26
Jenkins 2020 [74]	USA	English	Cervical surgery	202	48	50
Jorritsma 2012 [58]	Netherland	Dutch	Neck pain	112	63	39
Jorritsma 2012 [82]	Netherland	Dutch	Neck pain	76	71	38
Joseph 2015 [35]	India	Marathi	Neck pain	81	86	32
Jovicic 2018 [17]	Sebia	Serbian	Radiculopathy	50	70	45
Juul 2016 [94]	Denmark	Danish	Neck pain	196	75	48
Kaka 2016 [36]	Nigeria	Hausa	Neck pain	62	32	37
Langenfeld 2022 [95]	Switzerland	German	Neck pain	50	72	48
Lauridsen 2017 [51]	Denmark	Danish	Neck pain	326	63	45
Lim 2022 [37]	Singapore	Malay	Neck pain	120	86	45
Lim 2020 Singapore [38]	Singapore	Chinese	Neck pain	70	53	44

Lin 2020 [75]	China	Chinese	Cervical surgery	90	36	54
Lue 2018 [39]	Taiwan	Taiwanese	Neck pain	137	77	58
Luksanapruksa 2012 [52]	Thailand	Thai	Neck pain	46	74	44
Misterska 2011 [40]	Poland	Polish	Cervical surgery	60	57	47
Monticone 2015 [59]	Italy	Italian	Neck pain	200	60	53
Monticone 2012 [41]	Italy	Italian	Neck pain	101	66	48
Mousavi 2007 [42]	Iran	Persian	Neck pain	185	54	46
Nakamaru 2012 [53]	Japan	Japanese	Neck pain	110		60
Nieto 2008 [24]	Spain	Catalan	Whiplash	150	71	35
Odole 2011 [43]	Nigeria	Nigerian	Neck pain	32	66	48
Ortega 2010 [71]	Spain	Spanish	Neck pain	175	79	39
Owen 2019 [65]	USA	English	Cervical surgery	80	41	51
Owen 2018 [66]	USA	English	Cervical surgery	60	43	60
Papuga 2016 [67]	USA	English	Neck pain	283	48	55
Pennings 2020 [68]	USA	English	Cervical surgery	2018	48	57
Pereira 2015 [60]	Portugal	Portuguese	Neck pain	113	78	52
Pool 2007 [88]	Netherland	Dutch	Neck pain	183	61	46
Richardson 2012 [96]	USA	English	Cervical surgery	2070	56	43
Salehi 2019 [61]	Iran	Persian	Neck pain	57	77	38
Salo 2010 [77]	Finland	Finnish	Neck pain	101	58	50
Saltychev 2023 [84]	Finland	Finnish	Neck pain	338	51	54
Sandal 2021 [44]	India	Punjabi	Neck pain	115	62	37
Santiago-Reynoso 2021 [45]	Mexico	Mexican Spanish	Neck pain	113	66	30
Schuller 2014 [62]	Netherland	Dutch	Neck pain	1010		42
Shaheen 2013 [46]	Egypt	Arabic	Neck pain	65	31	41
Shashua 2016 [47]	Israel	Hebrew	Neck pain	100	61	53
Shrestha 2021 [18]	Nepal	Nepali	Radiculopathy	150	59	38
Song 2010 [48]	Korea	Korean	Cervical surgery	78	31	56
Swanenburg 2014 [27]	Switzerland	German	Mixed	49	73	39
Takeshita 2013 [63]	Japan	Japanese	Cervical surgery	130	32	59
Trouli 2008 [54]	Greece	Greek	Neck pain	68	55	62
Uthaikhup 2011 [49]	Thailand	Thai	Neck pain	185	74	49
Vaishnav 2020 [69]	USA	English	Cervical surgery	121	36	52
Van Der Velde 2009 [85]	Canada	English	Neck pain	521	65	45
Vernon 2010 [25]	Canada	English	Whiplash	107	50	45
Vos 2006 [26]	Netherland	Dutch	Neck pain (acute)	187	64	40
Walton 2013 [72]	Canada	English	Neck pain	316		44
Widbom-Kolhanen 2022 [55]	Finland	Finnish	Neck pain	392	52	55
Wu 2010 [50]	China	Chinese	Neck pain	125	62	43
Young 2009 [97]	USA	English	Neck pain	91	67	48
Young 2010 [19]	USA	English	Radiculopathy	165	65	49
Young 2019 [83]	USA	English	Neck pain	107	68	42

Table 2. Risk of systematic bias.

Study	1. Objective	2. Population description	3. Response rate >=50%	4. Inclusion criteria	5. Study power, variance reported	6. Outcome measures	TotaLow
Ailliet 2013	Low	Low	High	Low	High	Low	Low
Ailliet 2015	Low	Low	High	Low	Low	Low	Low
Aljinovic 2022	Low	Low	High	Low	Low	Low	Low
Aslan 2008	Low	Low	High	Low	Low	Low	Low
Bakhtadze 2021	Low	Low	High	Low	Low	Low	Low
Bakhtadze 2015	Low	Low	High	Low	Low	Low	Low
Bjorklund 2017	Low	Low	High	Low	High	Low	Low
Carreon 2010	Low	Low	High	High	High	Low	Low
Chien 2015	Low	Low	Low	High	High	High	High
Cleland 2008	Low	Low	Low	Low	Low	Low	Low
Cleland 2006	Low	Low	High	Low	Low	Low	Low
Cook 2006	Low	Low	High	Low	High	Low	Low
Cramer 2014	Low	Low	High	Low	Low	Low	Low
Croft 2016	Low	Low	High	Low	High	Low	Low
Cruz 2015	Low	Low	High	Low	Low	Low	Low
En 2009	Low	Low	High	Low	High	Low	Low
Farooq 2017	Low	Low	Low	Low	Low	U	High
Gabel 2016	Low	Low	High	High	High	Low	High
Gay 2007	Low	Low	Low	Low	High	Low	Low
Geete 2023	Low	Low	High	Low	Low	Low	Low
Geoghegan 2023	Low	Low	High	Low	High	Low	Low
Goh 2020	Low	Low	Low	Low	High	Low	Low
Hoving 2003	Low	Low	High	Low	High	Low	Low
Hung 2018	Low	Low	High	High	Low	Low	High
Hung 2019	Low	Low	High	High	Low	Low	High
Irmak 2019	Low	Low	High	Low	High	Low	Low
Jenkins 2020	Low	Low	Low	Low	High	Low	Low
Jorritsma 2012	Low	Low	High	Low	Low	Low	Low
Jorritsma 2012	Low	Low	High	Low	High	Low	Low
Joseph 2015	Low	Low	High	Low	High	Low	Low
Jovicic 2018	Low	Low	High	Low	High	Low	Low
Juul 2016	Low	Low	Low	Low	Low	Low	Low
Kaka 2016	Low	Low	High	Low	Low	Low	Low
Langenfeld 2022	Low	Low	High	Low	High	Low	Low
Lauridsen 2017	Low	Low	High	Low	U	Low	High
Lim 2022	Low	Low	Low	Low	Low	Low	Low

Lim 2020	Low	Low	Low	Low	Low	Low	Low
Lin 2020	Low	Low	High	Low	High	High	High
Lue 2018	Low	Low	High	Low	Low	Low	Low
Luksanapruksa 2012	Low	Low	High	Low	High	High	High
Misterska 2011	Low	Low	Low	Low	Low	Low	Low
Monticone 2015	Low	Low	Low	Low	High	High	Low
Monticone 2012	Low	Low	Low	Low	Low	Low	Low
Mousavi 2007	Low	Low	Low	Low	Low	Low	Low
Nakamaru 2012	Low	Low	Low	Low	Low	Low	Low
Nieto 2008	Low	Low	Low	Low	High	Low	Low
Odole 2011	Low	Low	Low	High	High	High	High
Ortega 2010	Low	Low	Low	Low	High	High	Low
Owen 2019	Low	Low	Low	Low	High	Low	Low
Owen 2018	Low	Low	Low	Low	Low	Low	Low
Papuga 2016	Low	Low	Low	Low	High	High	Low
Pennings 2020	Low	Low	Low	Low	Low	Low	Low
Pereira 2015	Low	Low	Low	High	Low	Low	Low
Pool 2007	Low	Low	Low	Low	Low	Low	Low
Richardson 2012	Low	Low	Low	Low	Low	Low	Low
Salehi 2019	Low	Low	Low	Low	Low	Low	Low
Salo 2010	Low	Low	Low	Low	Low	Low	Low
Saltychev 2023	Low	Low	Low	Low	Low	Low	Low
Sandal 2021	Low	Low	Low	Low	Low	Low	Low
Santiago-Reynoso 2021	Low	Low	Low	Low	Low	Low	Low
Schuller 2014	Low	Low	Low	Low	Low	Low	Low
Shaheen 2013	Low	Low	Low	Low	Low	Low	Low
Shashua 2016	Low	Low	Low	Low	Low	Low	Low
Shrestha 2021	Low	Low	Low	Low	Low	Low	Low
Song 2010	Low	Low	Low	High	Low	Low	Low
Swanenburg 2014	Low	Low	Low	Low	Low	Low	Low
Takeshita 2013	Low	Low	Low	Low	Low	Low	Low
Trouli 2008	Low	Low	Low	Low	Low	High	Low
Uthaikhup 2011	Low	Low	Low	Low	Low	Low	Low
Vaishnav 2020	Low	Low	Low	Low	Low	Low	Low
Van Der Velde 2009	Low	Low	Low	Low	Low	Low	Low
Vernon 2010	Low	Low	Low	Low	Low	Low	Low
Vos 2006	Low	Low	Low	Low	Low	Low	Low
Walton 2013	Low	Low	Low	Low	Low	Low	Low
Widbom-Kolhanen 2022	Low	Low	Low	Low	Low	Low	Low
Wu 2010	Low	Low	Low	Low	Low	Low	Low
Young 2009	Low	Low	Low	Low	Low	Low	Low
Young 2010	Low	Low	Low	Low	Low	High	Low
Young 2019	Low	Low	Low	Low	Low	Low	Low

Table 3. Psychometric characteristics reported by the included studies.

Study	Content validity	Face validity	Internal consistency	Factor structure	Test-retest reliability	Floor or ceiling effect	Pain VAS or NRS	Convergency with DASH	Convergency with PHQ	Convergency with SF-36	Convergency with PROMIS	Convergency with NPAD	Convergency with GROC	Convergent validity (other)	MDC or/and MCID	AUC	RASCH or IRT
Ailliet 2013	●			●				●									●
Ailliet 2015					●	●									●	●	
Aljinovic 2022					●	●	●		●								
Aslan 2008					●		●					●					
Bakhtadze 2021			●		●										●		
Bakhtadze 2015	●		●	●	●	●	●								●		
Bjorklund 2017															●	●	
Carreon 2010															●		
Chien 2015															●	●	
Cleland 2008					●										●	●	
Cleland 2006					●										●	●	
Cook 2006	●			●													
Cramer 2014			●	●	●		●										
Croft 2016				●													
Cruz 2015	●	●	●	●			●										
En 2009												●		●			
Farooq 2017	●	●	●	●	●	●	●						●	●	●		
Gabel 2016				●													
Gay 2007							●							●			
Geete 2023	●	●	●	●	●	●	●		●			●		●	●		
Geoghegan 2023							●			●							
Goh 2020																	
Hoving 2003														●			
Hung 2018															●		
Hung 2019											●						
Irmak 2019						●											
Jenkins 2020														●			
Jorritsma 2012			●			●						●					
Jorritsma 2012					●										●	●	
Joseph 2015	●	●	●		●												
Jovicic 2018		●	●	●		●	●		●					●			
Juul 2016			●		●										●		
Kaka 2016	●	●	●	●	●	●							●				
Langenfeld 2022																	
Lauridsen 2017		●	●	●													
Lim 2022	●	●	●	●	●	●	●								●		
Lim 2020	●	●	●	●	●	●	●						●		●		

Lin 2020														●			
Lue 2018	●	●	●	●	●	●	●							●			
Luksanapruksa 2012		●	●	●	●		●			●							
Misterska 2011	●		●		●	●								●			
Monticone 2015												●	●			●	
Monticone 2012	●	●	●	●	●	●	●			●		●		●			
Mousavi 2007	●	●	●		●	●	●					●					
Nakamaru 2012		●	●	●	●					●					●		
Nieto 2008			●	●													
Odole 2011	●	●			●												
Ortega 2010			●		●		●							●			
Owen 2019											●						
Owen 2018											●						
Papuga 2016											●						
Pennings 2020											●						
Pereira 2015						●							●		●	●	
Pool 2007															●		
Richardson 2012																	
Salehi 2019													●			●	
Salo 2010			●	●	●	●	●										
Saltychev 2023																	●
Sandal 2021	●	●	●	●	●	●	●										
Santiago-Reynoso 2021	●	●	●	●	●	●									●		
Schuller 2014													●		●		
Shaheen 2013	●	●	●	●	●	●											
Shashua 2016	●	●	●	●	●	●	●							●			
Shrestha 2021	●	●			●	●									●	●	
Song 2010	●	●	●	●	●		●			●							
Swanenburg 2014	●	●	●		●	●									●		
Takeshita 2013			●			●							●				
Trouli 2008		●	●	●	●	●									●		
Uthaikhup 2011	●	●	●	●	●	●	●										
Vaishnav 2020											●						
Van Der Velde 2009																	●
Vernon 2010							●							●			
Vos 2006					●	●									●		
Walton 2013					●		●							●	●		
Widbom-Kolhanen 2022			●	●													
Wu 2010	●	●	●		●							●					
Young 2009					●										●		
Young 2010					●										●	●	
Young 2019					●										●	●	

Table 4. Clinical relevancy of the results of the review.

NDI property	Results	Clinical significance
Content and face validity	Good for every studied translation	No direct significance to clinicians
Reliability (internal consistency)	Good Cronbach's alpha of 0.81 to 0.97	All 10 items are well correlated with each other, which supports the reliability of NDI composite score
Reliability (test-retest correlation)	Moderate to very high correlation of 0.56 to 0.98	NDI produces similar results when repeated
Reliability (test-retest between raters)	Excellent ICC of 0.91 (95% CI 0.90 to 0.93)	Same respondents probably end up with same results if fill NDI more than once
Construct validity	NDI will usually behave as a unidimensional score	NDI composite scores are comparable in most of situations
Convergent validity	NDI well correlates with such scales as NPAD and PROMIS, while correlations with pain severity, global improvement, SF-36 and PHQ-9 may significantly vary in different populations	Other common scales should be used along with the NDI to ensure the comprehensiveness of evaluation
MCID and MDC	MDC varied from 3% to 27% and MCID from 5% to 33%	It is safe to assume that change of 20%-35% in NDI score is the minimal change perceived by a respondent as meaningful
Discrimination ability	AUC was acceptable 0.74 (95% CI of 0.68 to 0.80)	NDI is well able to discriminate people who have improved from those who have not
Floor and ceiling effect	No floor or ceiling effect was detected by 25 studies.	NDI is able to produce reliable scores among those who have only mild disability as well as among people with severe disability
IRT and Rasch	Potential sex-related DIF for several items	NDI score may slightly depend on the sex of respondent

Supplement 1. Search strategy.

Database	Search clause
Medline	(“neck disability index”[TI] OR “ ndi “[TI]) AND (validity[TIAB] OR reliability[TIAB] OR “internal consistency” [TIAB] OR cronbach*[TIAB] OR alpha[TIAB] OR factor*[TIAB] OR “test-retest” [TIAB] OR “minimal clinically important” [TIAB] OR mcid[TIAB] OR “minimal detectable change” [TIAB] OR responsiveness[TIAB] OR “floor effect”[TIAB] OR “ceiling effect”[TIAB] OR “convergent validity” OR “criterion validity” OR “item response” [TIAB] OR rasch[TIAB] OR “differential item functioning”[TIAB]) Filters: Abstract, English
Embase	(“neck disability index”:ti OR “ndi”:ti) AND ((validity:ti OR reliability:ti OR consisten*:ti OR cronbach*:ti OR alpha:ti OR factor*:ti OR “test-retest”:ti OR “minimal clinically important”:ti OR mcid:ti OR “minimal detectable change”:ti OR responsiveness:ti OR floor:ti OR ceiling:ti OR “convergent validity” OR “criterion validity” OR “item response”:ti OR rasch:ti OR “differential item functioning”:ti) OR (validity:ab OR reliability:ab OR consisten*:ab OR cronbach*:ab OR alpha:ab OR factor*:ab OR “test-retest”:ab OR “minimal clinically important”:ab OR mcid:ab OR “minimal detectable change”:ab OR responsiveness:ab OR floor:ab OR ceiling:ab OR “convergent validity” OR “criterion validity” OR “item response”:ab OR rasch:ab OR “differential item functioning”:ab)) AND 'article'/it
Scopus	TITLE (“neck disability index” OR “ ndi ”) AND TITLE-ABS-KEY (validity OR reliability OR consisten* OR cronbach* OR alpha OR factor* OR “test-retest” OR “minimal clinically important” OR mcid OR “minimal detectable change” OR responsiveness OR floor OR ceiling OR “convergent validity” OR “criterion validity” OR “item response” OR rasch OR “differential item functioning”) AND LIMIT-TO (DOCTYPE, "ar")
Web of Science	TI=(“neck disability index”) AND (TI=(validity OR reliability OR consisten* OR cronbach* OR alpha OR factor* OR “test-retest” OR “minimal clinically important” OR mcid OR “minimal detectable change” OR responsiveness OR floor OR ceiling OR “convergent validity” OR “criterion validity” OR “item response” OR rasch OR “differential item functioning”) OR AB=(validity OR reliability OR consisten* OR cronbach* OR alpha OR factor* OR “test-retest” OR “minimal clinically important” OR mcid OR “minimal detectable change” OR responsiveness OR floor OR ceiling OR “convergent validity” OR “criterion validity” OR “item response” OR rasch OR “differential item functioning”))
Psycinfo	TI “neck disability index” Publication Type: Peer Reviewed Journal

Figure 1. Search flow (PRISMA)

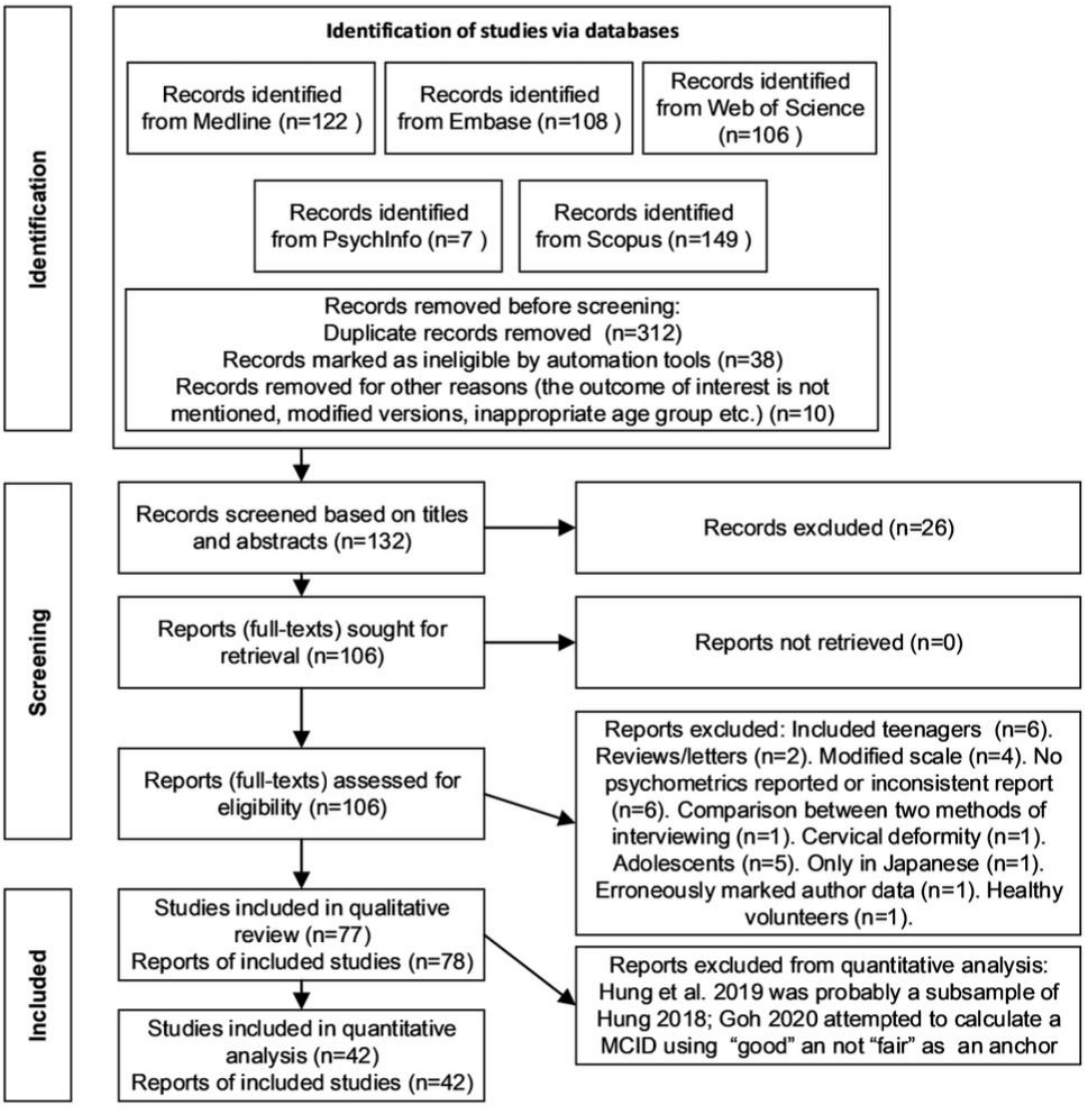


Figure 2. Forest plot of pooled intra-rater intraclass correlation (ICC) estimates.

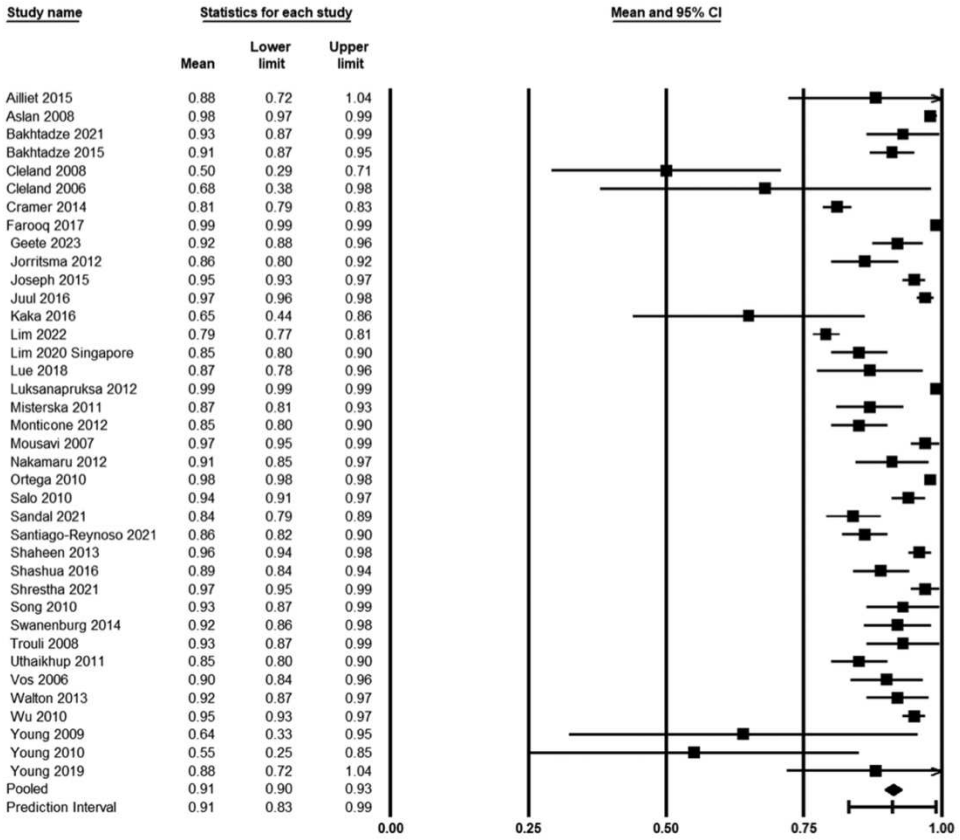


Figure 3. Funnel plot of publication bias for pooled test-retest intraclass correlation (ICC) and sensitivity (Area Under the Curve [AUC]).

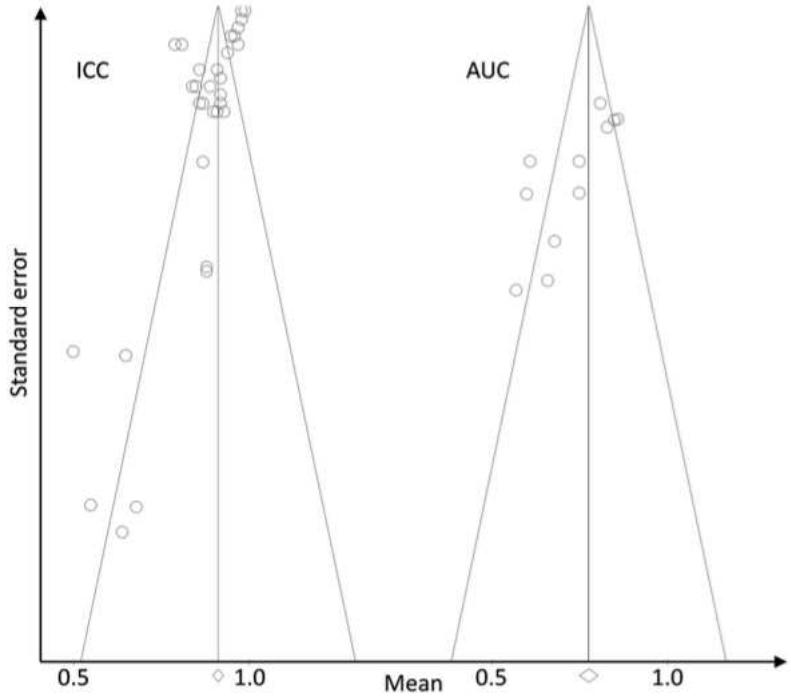
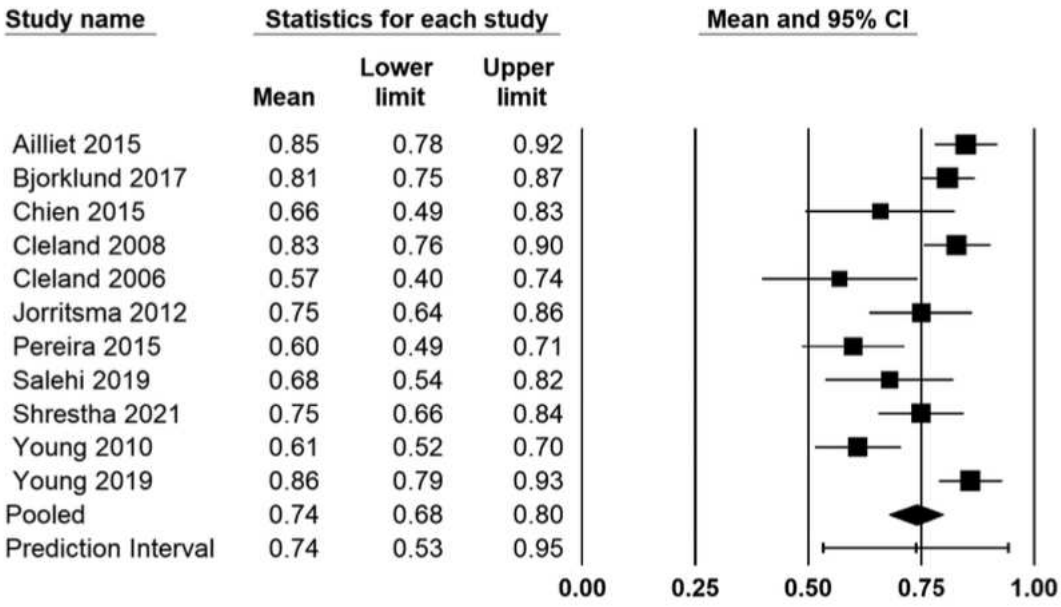


Figure 4. Forest plot of pooled Area Under the Curve (AUC) estimates.



6 REFERENCES

1. Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. Physiotherapy. 1980;66(8):271-3.
2. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther. 1991;14(7):409-15.
3. Vernon H. The Neck Disability Index: state-of-the-art, 1991-2008. J Manipulative Physiol Ther. 2008;31(7):491-502.
4. Vernon H. The psychometric properties of the Neck Disability Index. Arch Phys Med Rehabil. 2008;89(7):1414-5; author reply 5-6.
5. MacDermid JC, Walton DM, Avery S, Blanchard A, Etruw E, McAlpine C, et al. Measurement properties of the neck disability index: a systematic review. J Orthop Sports Phys Ther. 2009;39(5):400- 17.
6. Yao M, Sun YL, Cao ZY, Dun RL, Yang L, Zhang BM, et al. A systematic review of cross-cultural adaptation of the neck disability index. Spine (Phila Pa 1976). 2015;40(7):480-90.
7. Howell ER. The association between neck pain, the Neck Disability Index and cervical ranges of motion: a narrative review. J Can Chiropr Assoc. 2011;55(3):211-21.
8. Goh GS, Yue WM, Guo CM, Tan SB, Chen JL. Defining threshold values on the neck disability index corresponding to a patient acceptable symptom state in patients undergoing elective surgery for degenerative disorders of the cervical spine. Spine J. 2020;20(8):1316-26.
9. Hartman TJ, Nie JW, MacGregor KR, Oyetayo OO, Zheng E, Singh K. Neck Disability Index as a Prognostic Factor for Outcomes Following Cervical Disc Replacement. Clin Spine Surg. 2023;36(8):310- 6.
10. Toci GR, Lambrechts MJ, Karamian BA, Mao J, Heinle J, Bhatt S, et al. Depression Increases Posterior Cervical Decompression and Fusion Revision Rates and Diminishes Neck Disability Index Improvement. Spine (Phila Pa 1976). 2022;47(18):1287-94.
11. Hinkle D, Wiersma W, Jurs S. Applied Statistics for the Behavioral Sciences. 5th ed. Boston: Houghton Mifflin; 2003.
12. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276- 82.
13. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15(2):155-63.
14. Glen S. "Cronbach's Alpha: Definition, Interpretation, SPSS" From StatisticsHowTo.com: Elementary Statistics for the rest of us! 2023 [Available from: www.statisticshowto.com/probability- and-statistics/statistics-definitions/cronbachs-alpha-spss.
15. Mandrekar JN. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. J Thorac Oncol. 2010;5(9):1315-6.
16. 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(1):69-74.
17. Jovicic MD, Konstantinovic LM, Grgurevic AD, Milovanovic ND, Trajkovic G, Jovicic VZ, et al. Validation of the Neck Disability Index in Serbian Patients With Cervical Radiculopathy. J Manip Physiol Ther. 2018;41(6):496-502.
18. Shrestha D, Shrestha R, Grotle M, Nygaard Ø P, Solberg TK. Validation of the Nepali versions of the Neck Disability Index and the Numerical Rating Scale for Neck Pain. Spine (Phila Pa 1976). 2021;46(5):E325-e32.
19. Young IA, Cleland JA, Michener LA, Brown C. Reliability, construct validity, and responsiveness of the neck disability index, patient-specific functional scale, and numeric pain rating scale in patients with cervical radiculopathy. Am J Phys Med Rehabil. 2010;89(10):831-9.
20. Aljinović J, Barun B, Poljičanin A, Marinović I, Vlak T, Pivalica D, et al. Croatian version of the neck disability index can distinguish between acute, chronic and no neck pain: Results of a validation study. Wien Klin Wochenschr. 2022;134(3-4):162-8.
21. Croft AC, Milam B, Meylor J, Manning R. Confirmatory Factor Analysis and Multiple Linear Regression of the Neck Disability Index: Assessment If Subscales Are Equally Relevant in Whiplash and Nonspecific Neck Pain. J Chiropr Med. 2016;15(2):87-94.
22. Gabel CP, Cuesta-Vargas A, Barr S, Black SW, Osborne JW, Melloh M. Confirmatory factor analysis of the neck disability index, comparing patients with whiplash associated disorders to a control group with non-

specific neck pain. Eur Spine J. 2016;25(7):2078-86.
23. 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(3):273-81.
24. Nieto R, Miro J, Huguet A. Disability in subacute whiplash patients - Usefulness of the neck disability index. Spine. 2008;33(18):E630-E5.
25. Vernon H, Guerriero R, Kavanaugh S, Soave D, Moreton J. Psychological factors in the use of the neck disability index in chronic whiplash patients. Spine. 2010;35(1):E16-E21.
26. 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(11):1729-36.
27. Swanenburg J, Humphreys K, Langenfeld A, Brunner F, Wirth B. Validity and reliability of a German version of the Neck Disability Index (NDI-G). Man Ther. 2014;19(1):52-8.
28. Cook C, Richardson JK, Braga L, Menezes A, Soler X, Kume P, 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(14):1621-7.
29. Irmak R. Relatively short term test re-test reliability of Neck Disability Index by long term test re-retest reliability method of Oswestry Disability Index in healthy office workers. Work (Reading, Mass). 2019;64(3):635-40.
30. Ailliet L, Knol DL, Rubinstein SM, De Vet HCW, Van Tulder MW, Terwee CB. Definition of the construct to be measured is a prerequisite for the assessment of validity. the Neck Disability Index as an example. J Clin Epidemiol. 2013;66(7):775-82.
31. Bakhtadze MA, Vernon H, Zakharova OB, Kuzminov KO, Bolotov DA. The Neck Disability Index- Russian Language Version (NDI-RU): A Study of Validity and Reliability. Spine. 2015;40(14):1115-21.
32. Cruz EB, Fernandes R, Carnide F, Domingues L, Pereira M, Duarte S. Cross-cultural adaptation and validation of the neck disability index to european portuguese language. Spine. 2015;40(2):E77- E82.
33. Farooq MN, Mohseni-Bandpei MA, Gilani SA, Hafeez A. Urdu version of the neck disability index: A reliability and validity study. BMC Musculoskelet Disord. 2017;18(1).
34. Geete DB, Mhatre BS, Vernon H. Cross-cultural Adaptation and Psychometric Validation of the Hindi Version of the Neck Disability Index in Patients with Chronic Neck Pain. Spine (Phila Pa 1976). 2023.
35. Joseph SD, Bellare B, Vernon H. Cultural Adaptation, Reliability, and Validity of Neck Disability Index in Indian Rural Population. Spine. 2015;40(2):E68-E76.
36. Kaka B, Ogwumike OO, Vernon H, Adeniyi AF, Ogunlade SO. Cross-cultural adaptation, validity and reliability of the Hausa version of the Neck Disability Index questionnaire. Int J Ther Rehabil. 2016;23(8):380-5.
37. Lim HHR, Tan ST, Tang ZY, Yang M, Koh EYL, Koh KH. Cross-cultural adaptation and psychometric evaluation of the Malay version of the Neck Disability Index. Disabil Rehabil. 2022;44(1):124-30.
38. Lim HHR, Tang ZY, Hashim M, Yang M, Koh EYL, Koh KH. Cross-cultural Adaptation, Reliability, Validity, and Responsiveness of the Simplified-Chinese Version of Neck Disability Index. Spine (Phila Pa 1976). 2020;45(8):541-8.
39. Lue YJ, Chen CH, Chou SH, Lin CL, Cheng KI, Lu YM. Development and Validation of Taiwanese Version of the Neck Disability Index. Spine. 2018;43(11):E656-E63.
40. Misterska E, Jankowski R, Glowacki M. Cross-cultural adaptation of the Neck Disability Index and Copenhagen Neck Functional Disability Scale for patients with neck pain due to degenerative and discopathic disorders. Psychometric properties of the Polish versions. BMC Musculoskelet Disord. 2011;12.
41. Monticone M, Ferrante S, Vernon H, Rocca B, Dal Farra F, Foti C. Development of the Italian version of the neck disability index: Cross-cultural adaptation, factor analysis, reliability, validity, and sensitivity to change. Spine. 2012;37(17):E1038-E44.
42. Mousavi SJ, Parnianpour M, Montazeri A, Mehdian H, Karimi A, Abedi M, 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(26):E825-E31.
43. Odole AC, Adegoke BO, Akomas NC. Validity and test re-test reliability of the neck disability index in the

Nigerian clinical setting. Afr J Med Med Sci. 2011;40(2):135-8.
44. Sandal D, Jindal R, Gupta S, Garg SK, Vernon H. Reliability and Validity of Cross Culturally Adapted Punjabi Version of NDI (NDI-P) in Patients with Neck Pain: A Psychometric Analysis. Indian J Orthop. 2021;55(4):918-24.
45. Santiago-Reynoso GM, Alvarado-Luna AE, Fernandez-Matias R, Pecos-Martin D, Gallego- Izquierdo T. Transcultural adaptation of the neck disability index to mexican spanish and assessment of its psychometric properties. Eur Spine J. 2021;30(9):2654-60.
46. Shaheen AA, Omar MT, Vernon H. Cross-cultural adaptation, reliability, and validity of the Arabic version of neck disability index in patients with neck pain. Spine (Phila Pa 1976). 2013;38(10):E609-15.
47. Shashua A, Geva Y, Levran I. Translation, validation, and crosscultural adaptation of the Hebrew version of the neck disability index. Spine. 2016;41(12):1036-40.
48. Song KJ, Choi BW, Choi BR, Seo GB. Cross-cultural adaptation and validation of the korean version of the neck disability index. Spine. 2010;35(20):E1045-E9.
49. Uthaikhup S, Paungmali A, Pirunsan U. Validation of thai versions of the neck disability index and neck pain and disability scale in patients with neck pain. Spine. 2011;36(21):e1415-e21.
50. Wu S, Ma C, Mai M, Li G. Translation and validation study of Chinese versions of the neck disability index and the neck pain and disability scale. Spine. 2010;35(16):1575-9.
51. Lauridsen HH, O'Neill L, Kongsted A, Hartvigsen J. The Danish Neck Disability Index: New Insights into Factor Structure, Generalizability, and Responsiveness. Pain Pract. 2017;17(4):480-93.
52. Luksanapruksa P, Wathana-apisit T, Wanasinthop S, Sanpakit S, Chavasiri C. Reliability and validity study of a thai version of the neck disability index in patients with neck pain. J Med Assoc Thai. 2012;95(5):681-8.
53. Nakamaru K, Vernon H, Aizawa J, Koyama T, Nitta O. Crosscultural adaptation, reliability, and validity of the Japanese version of the Neck Disability Index. Spine. 2012;37(21):E1343-E7.
54. 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.
55. Widbom-Kolhanen S, Pernaa KI, Saltychev M. Reliability and validity of the Neck Disability Index among patients undergoing cervical surgery. International journal of rehabilitation research Internationale Zeitschrift fur Rehabilitationsforschung Revue internationale de recherches de readaptation. 2022;45(3):273-8.
56. Aslan E, Karaduman A, Yakut Y, Aras B, Simsek IE, Yaglý N. The cultural adaptation, reliability and validity of neck disability index in patients with neck pain: a Turkish version study. Spine (Phila Pa 1976). 2008;33(11):E362-5.
57. En MCC, 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. 2009;14(4):433-8.
58. Jorritsma W, De Vries GE, Dijkstra PU, Geertzen JHB, Reneman MF. Neck Pain and Disability Scale and Neck Disability Index: Validity of Dutch language versions. Eur Spine J. 2012;21(1):93-100.
59. Monticone M, Ambrosini E, Vernon H, Brunati R, Rocca B, Foti C, et al. Responsiveness and minimal important changes for the Neck Disability Index and the Neck Pain Disability Scale in Italian subjects with chronic neck pain. Eur Spine J. 2015;24(12):2821-7.
60. Pereira M, Cruz EB, Domingues L, Duarte S, Carnide F, Fernandes R. Responsiveness and Interpretability of the Portuguese Version of the Neck Disability Index in Patients With Chronic Neck Pain Undergoing Physiotherapy. Spine (Phila Pa 1976). 2015;40(22):E1180-6.
61. Salehi R, Negahban H, Saghayezhian N, Saadat M. The responsiveness of the persian version of neck disability index and functional rating index following physiotherapy intervention in people with chronic neck pain. Iran J Med Sci. 2019;44(5):390-6.
62. Schuller W, Ostelo RWJG, Janssen R, de Vet HCW. The influence of study population and definition of improvement on the smallest detectable change and the minimal important change of the neck disability index. Health Qual Life Outcomes. 2014;12(1).
63. Takeshita K, Hosono N, Kawaguchi Y, Hasegawa K, Isomura T, Oshima Y, et al. Validity, reliability and responsiveness of the Japanese version of the Neck Disability Index. J Orthop Sci. 2013;18(1):14- 21.

64. Hung M, Saltzman CL, Voss MW, Bounsanga J, Kendall R, Spiker R, et al. Responsiveness of the Patient-Reported Outcomes Measurement Information System (PROMIS), Neck Disability Index (NDI) and Oswestry Disability Index (ODI) instruments in patients with spinal disorders. Spine J. 2019;19(1):34-40.
65. Owen RJ, Khan AZ, McAnany SJ, Peters C, Zebala LP. PROMIS correlation with NDI and VAS measurements of physical function and pain in surgical patients with cervical disc herniations and radiculopathy. Journal of Neurosurgery: Spine. 2019;31(4):519-24.
66. Owen RJ, Zebala LP, Peters C, McAnany S. PROMIS physical function correlation with NDI and mJOA in the surgical cervical myelopathy patient population. Spine. 2018;43(8):550-5.
67. Papuga MO, Mesfin A, Molinari R, Rubery PT. Correlation of PROMIS physical function and pain CAT instruments with oswestry disability index and neck disability index in spine patients. Spine. 2016;41(14):1153-9.
68. Pennings JS, Khan I, Davidson CA, Freitag R, Bydon M, Asher AL, et al. Using PROMIS-29 to predict Neck Disability Index (NDI) scores using a national sample of cervical spine surgery patients. Spine J. 2020;20(8):1305-15.
69. Vaishnav AS, Gang CH, Iyer S, McAnany S, Albert T, Qureshi SA. Correlation between NDI, PROMIS and SF-12 in cervical spine surgery. Spine J. 2020;20(3):409-16.
70. Geoghegan CE, Mohan S, Lynch CP, Cha EDK, Jacob KC, Patel MR, et al. Validation of Neck Disability Index Severity among Patients Receiving One or Two-Level Anterior Cervical Surgery. Asian Spine J. 2023;17(1):86-95.
71. Ortega A, Martínez D, Ruiz A. Validation of the Spanish version of the Neck Disability Index. Spine (Phila Pa 1976). 2010;35(4):E114-8.
72. Walton DM, MacDermid JC. A brief 5-item version of the Neck Disability Index shows good psychometric properties. Health Qual Life Outcomes. 2013;11(1).
73. 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 Manip Physiol Ther. 2007;30(4):259-62.
74. Jenkins NW, Parrish JM, Nolte MT, Hrynewycz NM, Brundage TS, Singh K. Validating the VR-12 Physical Function Instrument After Anterior Cervical Discectomy and Fusion with SF-12, PROMIS, and NDI. Hss j. 2020;16(Suppl 2):443-51.
75. Lin T, Chen P, Wang Z, Chen G, Liu W. Does Cervical Sagittal Balance Affect the Preoperative Neck Disability Index in Patients with Cervical Myelopathy? Clin Spine Surg. 2020;33(1):E21-E5.
76. Cramer H, Lauche R, Langhorst J, Dobos GJ, Michalsen A. Validation of the German version of the neck disability index (NDI). BMC Musculoskelet Disord. 2014;15(1).
77. Salo P, Ylinen J, Kautiainen H, Arkela-Kautiainen M, Häkkinen A. Reliability and validity of the finnish version of the neck disability index and the modified neck pain and disability scale. Spine. 2010;35(5):552-6.
78. Gabel CP, Cuesta-Vargas AI, Osborne JW, Burkett B, Melloh M. Confirmatory factory analysis of the Neck Disability Index in a general problematic neck population indicates a one-factor model. Spine Journal. 2014;14(8):1410-6.
79. Bjorklund M, Wiitavaara B, Heiden M. Responsiveness and minimal important change for the ProFitMap-neck questionnaire and the Neck Disability Index in women with neck-shoulder pain. Qual Life Res. 2017;26(1):161-70.
80. Chien A, Lai DM, Cheng CH, Wang SF, Hsu WL, Wang JL. Responsiveness of the Chinese versions of the Japanese orthopaedic association cervical myelopathy evaluation questionnaire and neck disability index in postoperative patients with cervical spondylotic myelopathy. Spine. 2015;40(17):1315-21.
81. 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(5):598-602.
82. Jorritsma W, Dijkstra PU, De Vries GE, Geertzen JHB, Reneman MF. Detecting relevant changes and responsiveness of Neck Pain and Disability Scale and Neck Disability Index. Eur Spine J. 2012;21(12):2550-7.

83. Young IAPTD, Dunning JPD, Butts RPTP, Mourad FPD, Cleland JAPTP. Reliability, construct validity, and responsiveness of the neck disability index and numeric pain rating scale in patients with mechanical neck pain without upper extremity symptoms. Physiother Theory Pract. 2019;35(12):1328- 35.
84. Saltychev M, Widbom-Kolhanen SS, Pernaa KI. Sex-related differential item functioning of neck disability index. Disabil Rehabil. 2023:1-7.
85. 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 Care Res. 2009;61(4):544-51.
86. Roy JS, MacDermid JC, Woodhouse LJ. Measuring shoulder function: a systematic review of four questionnaires. Arthritis Rheum. 2009;61(5):623-32.
87. Costello AB, Osborne J. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis recommendations for getting the most from your analysis. Pract Assess Res Evaluation [Internet]. 2005 11.13.2023; 10(Article 7). Available from: https://scholarworks.umass.edu/pare/vol10/iss1/7
88. 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 (Phila Pa 1976). 2007;32(26):3047-51.
89. Ailliet L, Rubinstein SM, de Vet HCW, van Tulder MW, Terwee CB. Reliability, responsiveness and interpretability of the neck disability index-Dutch version in primary care. Eur Spine J. 2015;24(1):88-93.
90. Bakhtadze MA, Lusnikova IV, Bolotov DA, Kuzminov KO. Neck disability index in patients with cervicogenic headache. Russ J Pain. 2021;19(1):25-30.
91. Carreon LY, Glassman SD, Campbell MJ, Anderson PA. Neck Disability Index, short form-36 physical component summary, and pain scales for neck and arm pain: the minimum clinically important difference and substantial clinical benefit after cervical spine fusion. Spine Journal. 2010;10(6):469-74.
92. En MCC, 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. Manual Therapy. 2009;14(4):433-8.
93. Hung M, Saltzman CL, Kendall R, Bounsanga J, Voss MW, Lawrence B, et al. What are the MCIDs for PROMIS, NDI, and ODI instruments among patients with spinal conditions? Clin Orthop Relat Res. 2018;476(10):2027-36.
94. Juul T, Søgaard K, Davis AM, Roos EM. Psychometric properties of the Neck OutcOme Score, Neck Disability Index, and Short Form–36 were evaluated in patients with neck pain. J Clin Epidemiol. 2016;79:31-40.
95. Langenfeld A, Gassner AP, Wirth B, Mühlemann MB, Nyirö L, Bastiaenen C, et al. Responsiveness of the German version of the Neck Disability Index in chronic neck pain patients: a prospective cohort study with a seven-week follow-up. Arch Physiother. 2022;12(1):23.
96. Richardson SS, Berven S. The development of a model for translation of the Neck Disability Index to utility scores for cost-utility analysis in cervical disorders. Spine J. 2012;12(1):55-62.
97. Young BA, Walker MJ, Strunce JB, Boyles RE, Whitman JM, Childs JD. Responsiveness of the Neck Disability Index in patients with mechanical neck disorders. Spine J. 2009;9(10):802-8.