Pain and physical functioning in neuropathic pain: a systematic review of psychometric properties of various outcome measures
ABSTRACT

Introduction: A range of outcome measures across various domains are used to evaluate change following an intervention in clinical trials on chronic Neuropathic pain (NeP). However, in order to capture a real change in the variable of interest, the psychometric properties of a particular measure should demonstrate appropriate methodological quality. Various outcome measures in the domains of pain and physical functioning have been used in the literature for NeP, for which individual properties (e.g., reliability/validity) have been reported. To date, there is no definitive synthesis of evidence on the psychometric properties of those outcome measures, thus the aim of this systematic review was to evaluate the methodological quality [COnsensus based Standards for the selection of health status Measurement INstruments (COSMIN) guidelines] of studies that evaluated psychometric properties of pain and physical functioning outcome measures used for NeP.

Methods: Specific MeSH/key-words related to three areas (pain and/or physical functioning, psychometric properties, and NeP) were used to retrieve relevant studies (English language) in key electronic databases (Medline (Ovid), CINAHL (EBSCO), Scopus, AMED and Web of Science) from database inception- July 2012. Articles retrieval/screening and quality analysis (COSMIN) were carried out by two independent reviewers.

Results: 24 pain and 37 physical functioning outcome measures were identified, varying in methodological quality from Poor-Excellent.

Conclusion: Although a variety of pain and physical functioning outcome measures have been reported in the literature, few have demonstrate methodologically strong psychometric properties. Thus, future research is required to further investigate the psychometric properties of existing pain and physical functioning outcome measures used for clinical and research purposes.

Keywords: neuropathic pain; systematic review; pain; physical function; outcome measures; psychometric properties; reliability; validity; responsiveness
1. INTRODUCTION

Neuropathic pain (NeP) is defined by the International Association for the Study of Pain’s Neuropathic Pain Special Interest Group (NeuPSIG) as “pain arising as a direct consequence of a lesion or disease affecting the somatosensory system”. A range of assessment guidelines have been developed from the Initiative on Methods, Measurement and Pain Assessment in Clinical Trials (IMMPACT), the European Federation of Neurological Societies (EFNS), and the NeuPSIG for NeP clinical trials and for clinical practice. These guidelines advocate a range of measures for assessing the core domains of pain, quality of life, mood, sleep, and functional capacity (physical, cognitive, emotional, and social). This notwithstanding, a variety of outcome measures are available for the above stated domains. In order to evaluate the applicability of these measures, a systematic review of psychometric properties of available outcome measures used in published trials may provide a useful basis for selecting the best measurement instrument for a specific purpose.

Individual assessment of psychometric properties of available outcome measures is important. As part of this, in reviewing the evidence on available outcome measures, it is important to assess the methodological quality of those studies that investigated psychometric properties. While in clinical practice adoption of outcome measures will depend on feasibility of use (speed, ease of use, and limited need for an overly sophisticated instrument), emphases should be also be given to measures which are proven to be reliable, valid, and responsive/interpretable for a given population.

Pain remains a leading cause of disability at the individual level, associated with functional losses as well as mood disturbances. Thus the focus of this systematic review will be in evaluating the psychometric properties of various outcome measures used in the domains of pain and physical functioning in NeP. On examination of the literature, a number of outcome measures have been identified in which have been used to measure pain intensity and physical function in NeP trials; however, there is limited conclusive evidence on their psychometric properties. Use of reliable and valid outcome measures can help to better evaluate the patient’s outcomes in terms of
pain and physical functioning, enabling better management, including the earliest appropriate
management to minimize risks of co morbidities and disabilities.

Existing evidence on the psychometric properties of pain and physical functioning outcome
measures used in NeP trials have not previously been systematically reviewed. The aim of this
systematic review was to systematically review and identify the gaps in literature for the evaluated
psychometric properties (reliability, validity, responsiveness, and interpretability) of identified
outcome measures for ‘pain and physical functioning’ as recommended by the IMMPACT guidelines
in NeP population. This review involved a systematic search of the literature. The findings of the
current study may assist in outlining the effective intervention strategies for patients with NeP. The
objectives of this systematic review were:

• Systematically review and identify the type of established psychometric properties for the
  identified outcome measures quantifying pain and physical functioning in neuropathic pain
  populations.
• Evaluate the methodological quality of the included studies investigating the psychometric
  properties of the identified outcome measures in the domain of pain and physical
  functioning in neuropathic pain populations in accordance with the Consensus-based
  Standards for the selection of Health Measurement Instruments (COSMIN) checklist with 4-
  point scale.

2. METHOD

2.1 Information sources

A systematic search was conducted following the Preferred Reporting Items for Systematic
reviews and Meta-Analyses (PRISMA) guidelines. The following electronic databases were searched:
Ovid Medline, CINAHL, Scopus, AMED, and Web of Science (WOS) (from database inception to 31st
The search update engine from the available databases was activated in order to be familiar with the new searches in the current field, since the original search.

2.2 Search strategy

The key words and MESH headings in three broad areas (pain and/or physical functioning outcome measures, psychometric properties, and NeP) were used in the development of a search strategy (Table I). Several strategies were used to develop a comprehensive list of keywords/MeSH terms/subject headings representing each area. For outcome measures, all pain and physical functioning outcome measures that were used in clinical trials of NeP were chosen. For psychometric properties, we chose the standardised terminologies used by the COSMIN framework. For the terms relating to NeP, MESH terms/keywords indexed for neuropathy, neuralgia, and neurodynia were used. Words within each theme were combined with OR and across themes with AND. This search strategy was amended for different databases as necessary.

Insert Table I about here.

2.3 Study selection

Articles identified in the search underwent a series of screening processes. Firstly, duplicate articles were removed. Two reviewers (PM and LC) independently selected and screened articles for potential eligibility at the title and abstract stages. Full text articles of all potentially eligible abstracts were retrieved for application of the eligibility criteria. Disagreements between the reviewers regarding inclusion of individual studies were discussed during a consensus meeting and, when unresolved, were resolved by discussion with other reviewers (PH, CC, and GDB). References of the selected papers were further explored for relevant articles.

2.4 Eligibility criteria
Cross sectional studies and longitudinal cohort studies, which included at least one assessment of a psychometric property of a pain or functional outcome measure in a NeP population (Nep as defined by the Clinical Resource Efficiency Support Team- CREST) were included. The adopted search strategy revealed two distinct categories of evaluations: one intended for screening or diagnosis, and the other developed to measure outcomes. Since the focus of this review was to investigate the psychometric properties of tools used to measure changes in the status of either pain or functional outcomes over time: screening or diagnostic tools were excluded. Studies published as case report, editorial, or reviews were also excluded. Only articles published in the English language and on humans were selected.

2.5 Data extraction and synthesis

A systematic approach to data extraction was carried out by independent reviewers (PM and LC/ PH/ CC/ GDB), with equal number of articles randomly distributed among the team members. Each member extracted the data from the allotted articles, which were then checked for accuracy, with consensus meetings and opinions from other reviewers to resolve any disagreements. The following data were collected and tabulated from each of the included articles: study reference, participant characteristics, outcome measures studied, and type of psychometric properties tested (reliability and/or validity) (Table II). Further summary of identified outcome measures with their published psychometric properties and COSMIN grading were synthesized (Table IV & V). Results from excellent and good methodological quality studies based on COSMIN criteria (as stated in Table VI) were used to formulate recommendations for acceptable psychometric properties scores (for definitions of acceptable, good and excellent scores see Table VI).

2.6 Methodological quality of individual studies reporting on psychometric properties

Whereas a variety of tools are available to measure the methodological quality of studies that report on scale development and assessed psychometric properties, the Consensus-based Standards
for the selection of Health Measurement Instruments (COSMIN)\(^6\) checklist; developed by an international group of experts, is unique and preferred because it allows for individual assessment of each psychometric domain within a study.

The COSMIN checklist\(^{14}\) (Table III) consists of ‘A to J’ nine boxes (Internal consistency- Box A; Reliability- Box B; Measurement error- Box C; Content validity- Box D; Structural validity- Box E; Hypotheses testing- Box F; Cross-cultural validity- Box G; Criterion validity- Box H; Responsiveness-Box I; Interpretability- Box J), with 5–18 items concerning methodological standards for how each measurement property should be assessed. According to COSMIN guidelines, the methodological quality of a study is considered adequate if all items in a box (A to J) were considered adequate. For this, each item was scored on a 4-point rating scale (i.e., “poor”, “fair”, “good”, or “excellent”). The primary investigator (PM) independently scored all articles and the results were discussed and consensus obtained with each relevant team member. Methodological quality was determined using the ‘lowest rating score’\(^6\) achieved by any item for the representative psychometric property. Therefore, if one criterion for any property scored ‘poor’, the methodological quality for that particular property was rated as ‘poor’ overall, irrespective of the scores that other criteria achieved. Disagreements regarding COSMIN scoring were resolved by discussion between reviewers.

Reviewers were not blinded to the journal affiliation or authors of the included articles.

*Insert Table III about here.*

3. RESULTS

Figure 3.1 illustrates the study selection process. The search resulted in 10,913 articles. After accounting for duplicate removal, title screening, and abstract screening, 80 articles were identified and retrieved as potentially eligible for the review. While checking the eligibility of full text articles, a further 16 articles were excluded from the review as two articles were editorial papers; two were commentary papers; five articles were based on cancer pain; three papers were PhD publications;
and for the remaining four, full text article were not available. Thus total of 64 articles satisfied our eligibility criteria and were included in this review.

Insert Figure I about here

3.2 Characteristics of included studies

In total, 64 studies reporting 61 different outcome measures were identified. The included studies evaluated the psychometric properties of pain outcome domains (n=24) and physical function outcome domains (n=37), (Table II). For the 24 pain intensity outcome measures, fifteen (63%), measures were patient-reported/self-reported measures, and the rest nine (37%) were the therapist/clinician completed measures. For the 37 physical function outcome measures, seventeen (46%) measures were patient-reported/self-reported measures i.e. symptomatic assessment (subjective), nine (24%) measures were performance based measures, and the rest of the eleven (30%) measures were therapist completed measures i.e. symptoms and signs (subjective and objective testing). The synthesis of results per outcome measure, their published psychometric properties, and quality assessment scores for studies, are detailed in Table IV and V. Data on the characteristics of the study population and sample population were extracted on the interpretability and generalizability boxes provided by the COSMIN checklist. Information regarding the sample size and gender distribution is reported in Table II.

Insert Table II about here.

3.2.1 Pain intensity outcome measures

Pain domain outcomes (Table II, and IV) included: Brief Pain Inventory Scale for Diabetic Peripheral Neuropathy;\textsuperscript{15} Complex Regional Pain Syndrome Severity Score;\textsuperscript{16} Diabetes Symptom Checklist Type-2;\textsuperscript{17} Foot Function Index (pain subscale);\textsuperscript{18} Italian Neuropathic Pain Symptom Inventory;\textsuperscript{19} McGill Pain Questionnaire;\textsuperscript{20} modified Toronto Clinical Neuropathy Score;\textsuperscript{21} Neuropathic Pain Scale;\textsuperscript{22-24} Neuropathic Pain Sensory Inventory;\textsuperscript{25,26} 0-10 Numerical Rating Scale;\textsuperscript{27} Neuropathy
Total Symptom Score-6; 28 0-10 point Pain Intensity- Numerical Rating Scale; 29 Pain Quality Assessment Scale; 30,31 Portuguese version of the Neuropathic Pain Symptoms Inventory; 32
Quantitative Sensory Testing (hot and cold pain threshold); 33-35 Sensory evaluation with Semmens-Weinstein Monofilaments; 36 Short-form McGill Pain Questionnaire-2; 37 Spanish Neuropathic Pain Symptom Inventory; 38 Toronto Clinical Scoring System; 39 Total Neuropathy Score; 40 Trauma Related Neuronal Dysfunction Symptoms Inventory; 41 Utah Early Neuropathy Scale; 42 Visual Analog Scale; 43 and Zoster Brief Pain Inventory. 44,45

3.2.2 Physical functioning outcome measures

The range of physical functioning outcome measures was equally extensive, and included (Table II, and V): Alderson-McGall Hand Function questionnaire; 46 Barthel Index; 47 Berg Balance Measure; 48 Brief Pain Inventory Facial; 49 Charcot-Marie-Tooth disease Neuropathy score; 50,51 Charcot-Marie-Tooth disease Neuropathy Score-2; 52 Disabilities of Arm, Shoulder and Hand Questionnaire; 53-56 Deambulation Index; 47 Dellon-modified Moberg pick-up test; 57 Facial Disability Index; 58 Functional Dexterity test; 59 Human Activity Profile; 60 INCAT The Overall Disability Sum Score; 61 Inflammatory neuropathy Sensory Score; 62 Levine-Katz Questionnaire; 56 Michigan Hand Outcome Questionnaire; 53 modified Neuropathy Disability Score; 63 10-Meter walking test; 48,64 Nine-Hole Peg test; 64 Neuropathy Impairment Score; 51 Overall Disability Sum Score; 65 Overall Neuropathy Limitations Scale; 64,66 Patient Evaluation Measure; 53 Physical Performance Measures (6 minute walk test, Timed up and go test); 67 Questionnaire Rising and Sitting down; 68 Radboud skills Questionnaire; 69 short form Screening of Activity Limitation and Safety Awareness Scale; 70,71 Step Activity Monitor; 72 Step Activity Monitor (4 min walk test); 73 Sheehan Disability Scale; 74 Sollerman Hand function test; 59 Turkish version of the Boston Questionnaire; 75 Ulnar Neuropathy at the Elbow Questionnaire; 76 12-Item Multiple Sclerosis Walking Scale; 77 Walking Stairs Questionnaire; 68 Work stimulation tasks (knob turn, Linear motion, and Lever arm); 78 and Zoster Impact Questionnaire. 45
3.3 Methodological quality of studies evaluating psychometric properties of pain intensity and physical functioning outcome measures

3.3.1 Reliability

The majority of the instruments included in our review were not tested for all psychometric properties listed on COSMIN checklist. Forty four of the sixty four studies (68%) assessed various forms of reliability (Internal consistency, inter-rater reliability, intra-rater reliability, test-retest reliability, and measurement error) and showed a mixed methodological quality of evidence (excellent/good/fair/poor), when evaluated on COSMIN (Table IV and V). The key results for reliability showed that the BPI-DPN, and the SF-MPQ2 have excellent ($\alpha>0.90$) internal consistency. The mTCNS has good internal consistency ($\alpha$= 0.81-0.90), inter-rater reliability, and intra-rater reliability (ICC or $K=0.81-0.90$). The hot and cold pain thresholds on the QST have good inter-rater and test-retest reliability (ICC or $K=0.81-0.90$). The Spanish NPSI has excellent internal consistency($\alpha>0.90$) with good test-retest reliability(ICC or $K=0.81-0.90$). Measurement error was the least reported form of reliability, and the TRNDSI had good test-retest reliability (ICC or $K=0.81-0.90$) and measurement error (see Table IV). These measures with excellent and good psychometric properties scores also scored good/excellent on the COSMIN checklist (as according to COSMIN criteria stated in Table VI).

3.3.2 Validity

Validity was the more frequently tested psychometric property, in forty nine of sixty four studies (76%), there was face/content validity, structural validity, construct validity, criterion/concurrent validity, convergent validity, discriminative validity, hypothesis testing, and responsiveness. Similar to the findings for reliability, mixed methodological quality evidence (excellent/good/fair/poor) was found when evaluated on COSMIN (Table IV and V). The key results for validity showed that the NPSI, the SALSA, and the UNEQ have excellent content validity as there
were no concerns raised by the patients or experts regarding the wording of questionnaires, and thus no further modifications were advised. The UENS has the best criterion validity followed by the HAP and the mNDS. Approximately one third of the studies (18/49, 36%) evaluated responsiveness form of validity. The NPS has excellent responsiveness followed by the 0-10 PI NRS, and the ODSS. Also the studies showing these evidences were of excellent/good methodological quality on the COSMIN checklist [as according to COSMIN criteria stated in Table VI].

Insert Table IV and V about here.

4. Discussion

To our knowledge, this is the first systematic review to evaluate the evidence for the psychometric properties of pain and physical functional outcome measures used in assessment in NeP conditions, and to identify the methodological quality of the studies investigating the psychometric properties of various outcome measures. A total of 61 different outcome measures were identified related to the domains of pain and physical functioning. In this systematic review, while most of the studies have shown good/excellent evidence of reliability and validity of the used scales, only few are considered ‘excellent to good’ in terms of their methodological quality. Our review identified acceptable reliability and validity (for a few key properties) for the mTCNS, the TRNDI, the 0-10 PI NPS, the QST, the SALSA, the Spanish NPSI, the ODSS, the SF-MPQL, the UNEQ, the UENS, the HAP, the mNDS, the NDS and the BPI-DPN.

The available studies investigating the psychometric property of reliability were rated in varying methodological quality from ‘poor’ to ‘excellent’ on the COSMIN checklist. However, the majority of studies showed similar methodological shortcomings. In this review, smaller sample sizes were found to be associated with the majority of inconsistent results. According to COSMIN guidelines, a sample size of ≥100 is considered to be an adequate/excellent sample size, given the need for precision in the overall estimates; these estimates are based on the power 0.80.
A sample size of 50 provides a 0.70 power (level of significance being 0.05), while 100 has a power of 0.94.\textsuperscript{25}

In the current systematic review, many outcome measures seem promising for different domains of reliability and validity (according to COSMIN criteria stated in Table VI), as the FFI, the NTSS-6, the AMHFQ, the DASH, the HAP, the ISS, the MHQ, the PEM, the SDS, the TBQ, the UNEQ, and the Walk-12 scales have ‘moderate’ ($\alpha>0.71-0.80$) to ‘excellent’ ($\alpha>0.90$) published grades for internal consistency. However, when the methodological quality of the studies were evaluated on COSMIN, these were graded of ‘poor/fair’ quality because of the small sample size. These findings are consistent with those of a recent systematic review on outcome measures in neck pain, where smaller sample sizes frequently led to poorer results.\textsuperscript{80} This current review recommends that future research on a larger sample size ($n=\geq100$, as recommended by COSMIN) is needed to improve the quality of research on these measures.

Validity was the most frequently evaluated psychometric property in both pain and physical functioning outcome domains. The majority of these studies demonstrated unsatisfactory (poor/fair scores) results on COSMIN. The main reasons for this were inconsistencies in the following areas: smaller sample sizes; hypotheses were not formulated; and expected direction/magnitude of correlations was not stated in advance. Other common findings were a lack of information about reporting of missing items, and measures adopted to handle missing data. Though these two items did not contribute to the overall ‘poor’ grading on the COSMIN, it is expected that studies of ‘good’ methodological quality should report this construct, as a high number of missing items can introduce bias.

A further interesting finding of this review was that responsiveness was the least frequently studied psychometric property for the included pain and physical functioning outcome measures. There were a total of 18 studies which published the findings on responsiveness and only three scales- the NPS, the 0-10 PI NRS and the ODSS proved satisfactory methodological quality on
COSMIN. The remaining measures were graded ‘fair to poor’, and all the above stated shortcomings (small sample size, un-reporting of missing items, vagueness about how the missing data were handled, not well formulated hypothesis etc.) equally contributed to the inconsistent results for the studies reporting on this property.

In the current systematic review, there were few measures identified which had promising psychometric properties for key variables: the mTCNS (good internal consistency, inter-rater and intra-rater reliability and criterion validity); the TRNDISI, and the ZBPI (good test-retest reliability); the NPSI (excellent face/content validity); the 0 to 10 PI NRS (good responsiveness); the QST- pain threshold (good intra-rater and test-retest reliability); the NPS (excellent responsiveness); and the SALSA (excellent internal consistency and content validity), and were supported by a “excellent to good’ methodological quality on the COSMIN checklist. The future use of these measures can be recommended based on their proven psychometric properties; however, it is imperative that other remaining psychometric properties of these outcome measures should also be established.

We also identified a list of instruments which showed their best methodological quality for few psychometric properties on COSMIN, but at the same time good methodological quality evidence was lacking for other properties: the TCSS (good construct validity, but poor inter and intra-rater reliability); the Short-form MPQ- 2 (excellent internal consistency, but fair construct validity and responsiveness); the HAP (good criterion validity, with poor internal consistency and responsiveness and fair hypothesis testing); the ODSS (good responsiveness but fair inter-rater and intra-rater reliability and construct validity); the UNEQ (excellent content validity, fair test-retest reliability, and poor internal consistency, construct validity, and responsiveness); the TBQ (good construct validity, fair test-retest reliability, and poor internal consistency); the UENS (excellent criterion validity, with poor inter-rater reliability and responsiveness); and the BPI-DPN (excellent internal consistency and discriminative validity, fair construct validity and poor criterion validity).

Since study methodology may influence results for psychometric properties, it is recommended that
further evaluation of these psychometric properties with studies of improved methodological quality should be carried out.

Limitations

Firstly, it is acknowledged that ‘Neuropathic Pain conditions’ is an umbrella term which covers a range of different conditions such as diabetic neuropathy, trigeminal neuralgia, and post herpetic neuralgia. For the search strategy, MESH terms/ key words indexed for neuropathy, neuralgia, and neurodynia were used to be as inclusive as possible. It is acknowledged that each condition could have been separately searched, and that such an approach may have lessened the chances of missing studies.

Secondly, psychometric properties such as reliability and validity, including responsiveness, are sub classified into various forms such as internal consistency, inter-rater/test retest reliability, content validity, minimal important difference, and standard error of measurement etc. For the current search strategy, keywords in three broader areas (reliability and/or, validity and/or, and responsiveness) were used rather than individual sub classified keywords. However, since these broader terms are the most commonly used to denote the various forms of psychometric properties, it is anticipated that the majority of studies would have been selected.

Lastly, for this systematic review, multidisciplinary, international consensus-based methodological quality reporting guidelines, COSMIN, were followed for rating the quality of included studies of psychometric properties. The COSMIN checklist has well developed data extraction forms with detailed instructions for completion. The 4-point rating scale classifies each assessment of a measurement property as ‘excellent, good, fair, or poor’, based on the scores of the items in the corresponding COSMIN box. The methodological quality of a study is considered adequate if all items in a box (A to J) are considered adequate. However, frequently not all items in a box are scored adequate, and it is not feasible to provide overall definitive grade for each
psychometric property; thus no decisions can be drawn for the methodological quality of the studies based purely on COSMIN findings.

Conclusion

In this review we evaluated the evidence for psychometric properties of 61 unique outcome measures identified to assess pain and physical functioning outcome domains in trials of NeP conditions. We have presented extensive data which demonstrate the psychometric properties of these available outcome measures, and recommend the use of the mTCNS, the TRNDSI, the ZBPI, the NPSI, the 0 to 10 PI NRS, the QST- pain threshold, and the NPS to detect changes in pain intensity and physical functions. We found that important information regarding the methodological quality of the majority of studies demonstrating these psychometric properties is lacking or is of poor quality. Since NeP is a multi-disabling condition with significant associated morbidity, usage of quality evidenced pain and physical functional measures is a key recommendation for future research in NeP intervention studies. It appears that despite representing these measures in many studies of NeP, the methodological quality for most of the measures is not strong enough to recommend their use based on their psychometric properties. Thus, good quality future research is required to further investigate the psychometric properties of identified outcome measures used for clinical and research purposes.

ACKNOWLEDGEMENTS

All authors declare that there exist no conflicts of interest associated with the current study. The authors thank and acknowledge Dr Ramakrishnan Mani, Lecturer, Centre for Health Activity and Rehabilitation Research, School of Physiotherapy, University of Otago for his suggestions, invaluable constant assistance and helping with the constructive feedback on drafts of the manuscript. The findings of the study have been presented as a poster in the 8th Congress of the European Federation of IASP® Chapters (EFIC 2013) in Florence, Italy.
REFERENCES


27. Farrar JT, Pritchett YL, Robinson M, Prakash A, Chappell A. The clinical importance of changes in the 0 to 10 numeric rating scale for worst, least, and average pain intensity:


**FIGURE LEGENDS**

**Figure I** Flow diagram summarising study selection process
<table>
<thead>
<tr>
<th>Theme 1</th>
<th>AND</th>
<th>Theme 2</th>
<th>AND</th>
<th>Theme 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychometric properties</td>
<td></td>
<td>Pain and/or Physical functional Outcome measures</td>
<td></td>
<td>Neuropathic Pain</td>
</tr>
<tr>
<td>Clinometric properties OR</td>
<td></td>
<td>Visual Analog Scale OR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validity OR</td>
<td></td>
<td>Numerical Pain Rating scale OR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reliability OR</td>
<td></td>
<td>McGill pain rating scale OR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity OR</td>
<td></td>
<td>Pain disability index OR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Responsiveness OR</td>
<td></td>
<td>Functional component of The Western Ontario and McMaster Universities Arthritis Index OR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimal(ly) clinically important difference OR</td>
<td></td>
<td>Timed scored functional activity OR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimal(ly) clinically important change OR</td>
<td></td>
<td>Functional reach test OR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minimum detectable change OR</td>
<td></td>
<td>Timed 9.1 metre to walk OR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smallest detectable change</td>
<td></td>
<td>Disability of the arm shoulder and hand</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>questionnaire OR</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ulnar Neuropathy at Elbow questionnaire OR</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Daily activities by Verbal Rating Scale OR</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Function interference by Numerical Rating Scale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Participant’s characteristics</td>
<td>Outcome measures studied</td>
<td>Psychometric properties tested</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-------------------------------</td>
<td>--------------------------</td>
<td>-------------------------------</td>
<td></td>
</tr>
<tr>
<td>Alderson &amp; McGall 1999</td>
<td>Carpal Tunnel Syndrome n= 17 Gender = 5 M, 12 F</td>
<td>Alderson-McGall hand function questionnaire</td>
<td>Reliability- Internal consistency, test-retest reliability; Validity- Convergent validity</td>
<td></td>
</tr>
<tr>
<td>Amirjani et al. 2011</td>
<td>Carpal Tunnel Syndrome n= 162 Gender = 120 M, 42 F</td>
<td>Dellon-modified Moberg pick-up test</td>
<td>Reliability- test-retest reliability; Validity- Discriminative validity</td>
<td></td>
</tr>
<tr>
<td>Asad et al. 2010</td>
<td>Type 2 diabetics n= 60 sensorimotor NeP Gender = not mentioned</td>
<td>modified Neuropathy Disability Score</td>
<td>Validity- Criterion validity</td>
<td></td>
</tr>
<tr>
<td>Bastyr et al. 2005</td>
<td>Diabetic peripheral NeP n= 205 Gender = 122 M, 83 F</td>
<td>Neuropathy Total Symptom Score-6</td>
<td>Reliability- Internal consistency, test-retest reliability; Validity- Construct &amp; Convergent validity, Responsiveness</td>
<td></td>
</tr>
<tr>
<td>Bouhassira et al. 2004</td>
<td>Peripheral and Central NeP n= 176 Gender = 97 M, 79 F</td>
<td>Neuropathic Pain Symptom Inventory</td>
<td>Reliability- test-retest reliability; Validity- Face validity, Structural validity, Criterion validity, Convergent validity, Divergent validity, Responsiveness</td>
<td></td>
</tr>
<tr>
<td>Study Reference</td>
<td>Condition</td>
<td>Sample Size</td>
<td>Gender Distribution</td>
<td>Scoring System/Measurements</td>
</tr>
<tr>
<td>-----------------------</td>
<td>---------------------------------------</td>
<td>-------------</td>
<td>---------------------</td>
<td>-------------------------------------------------------------------</td>
</tr>
<tr>
<td>Bril &amp; Perkins 2002</td>
<td>Type 1 and 2 diabetic NeP</td>
<td>n = 89</td>
<td>65 M, 24 F</td>
<td>Toronto Clinical Scoring System</td>
</tr>
<tr>
<td>Bril et al. 2009</td>
<td>Diabetic sensorimotor poly NeP</td>
<td>n = 65</td>
<td>40 M, 25 F</td>
<td>modified Toronto Clinical Neuropathy Score</td>
</tr>
<tr>
<td>Collins et al. 2008</td>
<td>Complex regional pain syndrome-I</td>
<td>n = 27</td>
<td>5 M, 22 F</td>
<td>Trauma Related Neuronal Dysfunction Symptoms Inventory</td>
</tr>
<tr>
<td>Coplan et al. 2004</td>
<td>Herpes Zoster</td>
<td>n = 121</td>
<td>45 M, 76 F</td>
<td>Zoster Brief Pain Inventory Questionnaire</td>
</tr>
<tr>
<td>Cornblath et al. 1999</td>
<td>Diabetic poly NeP</td>
<td>n = 30</td>
<td>18 M, 12 F</td>
<td>Total Neuropathy Score</td>
</tr>
<tr>
<td>Crawford et al. 2008</td>
<td>Neuropathic Pain</td>
<td>n = 130</td>
<td>70 M, 60 F</td>
<td>Neuropathic Pain Symptom Inventory questionnaire</td>
</tr>
<tr>
<td>Davidoff et al. 1988</td>
<td>Reflex Sympathetic Dystrophy Syndrome</td>
<td>n = 17</td>
<td>5 M, 12 F</td>
<td>Visual Analog Scales</td>
</tr>
<tr>
<td>de Andrade et al. 2011</td>
<td>Neuropathic Pain</td>
<td>n = 94</td>
<td>57 M, 37 F</td>
<td>Portuguese Neuropathic Pain Symptoms Inventory</td>
</tr>
<tr>
<td>Study</td>
<td>Population Description</td>
<td>Sample Size</td>
<td>Instruments</td>
<td>Reliability</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-----------------------------------------------</td>
<td>-------------</td>
<td>----------------------------------------------------------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Dias et al. 2008</td>
<td>Wrist and hand disorders due to nerve involvement</td>
<td>n= 26</td>
<td>The Patient Evaluation Measure; The Michigan Hand Outcome Questionnaire; The Disabilities of Arm, Shoulder and Hand Questionnaire</td>
<td>Internal consistency, test-retest</td>
</tr>
<tr>
<td>Dworkin et al. 2009</td>
<td>Diverse chronic pain syndrome; Diabetic NeP</td>
<td>n= 1108</td>
<td>Short-form McGill Pain Questionnaire- 2</td>
<td>Internal consistency; Construct validity, Responsiveness</td>
</tr>
<tr>
<td>Eklund et al. 2009</td>
<td>Charcot-Marie-Tooth disease</td>
<td>n= 20</td>
<td>The Disabilities of Arm, Shoulder and Hand Questionnaire</td>
<td></td>
</tr>
<tr>
<td>Erdmann et al. 2005</td>
<td>Chronic idiopathic demyelinating polyneuropathy; Multifocal Mono neuropathy</td>
<td>n= 30</td>
<td>Berg Balance Measure; 10 meter walk test</td>
<td></td>
</tr>
<tr>
<td>Farrar et al. 2010</td>
<td>Diabetic peripheral NeP; Fibromyalgia syndrome</td>
<td>n= 1700</td>
<td>0 to 10 Numeric Rating Scale</td>
<td></td>
</tr>
<tr>
<td>Farrar et al. 2001</td>
<td>Diabetic peripheral NeP</td>
<td>n= 984</td>
<td>0 to 10 point Pain Intensity</td>
<td></td>
</tr>
<tr>
<td>Authors</td>
<td>Condition</td>
<td>n</td>
<td>Gender</td>
<td>Measured Pain Tools</td>
</tr>
<tr>
<td>------------------</td>
<td>----------------------------------</td>
<td>-----</td>
<td>--------</td>
<td>-------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Farrell et al.</td>
<td>Post Herpetic Neuralgia</td>
<td>31</td>
<td>56 M, 41 F</td>
<td>Numerical Rating Scale</td>
</tr>
<tr>
<td>Felix &amp; Widerstrom-Noga</td>
<td>NeP related to Spinal Cord Injury</td>
<td>22</td>
<td>19 M, 3 F</td>
<td>Human Activity Profile</td>
</tr>
<tr>
<td>Galer &amp; Jensen</td>
<td>Post Herpetic Neuralgia; Diabetic NeP; Peripheral Nerve Injury</td>
<td>160 (69; 24; 67)</td>
<td>not mentioned</td>
<td>Quantitative Sensory Testing (cold and heat pain thresholds)</td>
</tr>
<tr>
<td>Geber et al. 2011</td>
<td>Peripheral Nerve lesion; Other neuropathies</td>
<td>60</td>
<td>37 M, 23 F</td>
<td>The Neuropathic Pain Scale</td>
</tr>
<tr>
<td>Graham &amp; Hughes</td>
<td>Peripheral NeP</td>
<td>65</td>
<td>36 M, 29 F</td>
<td>12-Item Multiple Sclerosis Walking Scale</td>
</tr>
<tr>
<td>Graham &amp; Hughes</td>
<td>Peripheral NeP</td>
<td>100</td>
<td>51:49</td>
<td>The Overall Neuropathy Limitations Scale</td>
</tr>
<tr>
<td>Study</td>
<td>Condition</td>
<td>Sample Size</td>
<td>Gender</td>
<td>Measure/Scale</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------------------------------------------</td>
<td>-------------</td>
<td>-----------------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>Harden et al. 2010</td>
<td>Complex and non-complex regional pain syndrome</td>
<td>n= 155</td>
<td>Gender = 68 M, 87 F</td>
<td>Complex regional pain syndrome severity score</td>
</tr>
<tr>
<td>Helme et al. 1989</td>
<td>Chronic Neuropathic Pain due to Post Herpetic Neuralgia</td>
<td>n= 49</td>
<td>Gender = 10 M, 39 F</td>
<td>McGill Pain Questionnaire</td>
</tr>
<tr>
<td>Jensen et al. 2005</td>
<td>Peripheral NeP</td>
<td>n= 133</td>
<td>Gender = 63 M, 70 F</td>
<td>The Neuropathic Pain Scale</td>
</tr>
<tr>
<td>Jensen et al. 2006</td>
<td>Diabetes related foot pain</td>
<td>n= 159</td>
<td>Gender = 83 M, 76 F</td>
<td>The Neuropathic Pain Scale</td>
</tr>
<tr>
<td>Jensen et al. 2006</td>
<td>Carpal Tunnel Syndrome</td>
<td>n= 40</td>
<td>Gender = 12 M, 2 F</td>
<td>Pain Quality Assessment Scale</td>
</tr>
<tr>
<td>Jensen et al. 2010</td>
<td>Carpal Tunnel Syndrome</td>
<td>n= 100</td>
<td>Gender = 75 M, 25 F</td>
<td>Pain Quality Assessment Scale</td>
</tr>
<tr>
<td>Kilmer et al. 2000</td>
<td>Hereditary motor and sensory NeP</td>
<td>n= 9</td>
<td>Gender = 3 M, 6 F</td>
<td>Work stimulation tasks; Hand-held dynamometry</td>
</tr>
<tr>
<td>Study</td>
<td>Condition</td>
<td>Sample Size</td>
<td>Gender Distribution</td>
<td>Measure(s)</td>
</tr>
<tr>
<td>------------------</td>
<td>----------------------------------</td>
<td>-------------</td>
<td>---------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Lee et al. 2010</td>
<td>Typical &amp; atypical facial pain due to Trigeminal Neuralgia</td>
<td>n= 156</td>
<td>Gender = 58 M, 98 F</td>
<td>Brief Pain Inventory- Facial</td>
</tr>
<tr>
<td>Manor et al. 2008</td>
<td>Peripheral NeP</td>
<td>n= 20</td>
<td>Gender = 8 M, 12 F</td>
<td>Physical Performance Measures</td>
</tr>
<tr>
<td>Maser et al. 1989</td>
<td>Diabetic neuropathy</td>
<td>n= 100</td>
<td>Gender = 54 M, 46 F</td>
<td>Quantitative sensory testing (thermal sensitivity)</td>
</tr>
<tr>
<td>Melchior &amp; Velema</td>
<td>Leprosy related Neuropathic Pain</td>
<td>n= 25</td>
<td>Gender = not mentioned</td>
<td>Screening of Activity Limitation and Safety Awareness Scale</td>
</tr>
<tr>
<td>Merkies &amp; Schmitz</td>
<td>Guillain Barré Syndrome; Chronic idiopathic demyelinating neuropathy</td>
<td>n= 20</td>
<td>Gender = 12 M, 8 F</td>
<td>The INCAT Overall Disability Sum Score</td>
</tr>
<tr>
<td>Merkies et al. 2002</td>
<td>Neuropathic Pain</td>
<td>n= 113</td>
<td>Gender = not mentioned</td>
<td>The Overall Disability Sum Score</td>
</tr>
<tr>
<td>Reference</td>
<td>Study Type</td>
<td>Sample Size</td>
<td>Measurement</td>
<td>Reliability</td>
</tr>
<tr>
<td>--------------------</td>
<td>-----------------------------</td>
<td>-------------</td>
<td>--------------------------------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>Merkies et al. 2000</td>
<td>Neuropathic Pain</td>
<td>n= 113</td>
<td>Inflammatory Sensory Score</td>
<td>Reliability- Internal consistency, inter-rater, intra-rater reliability;</td>
</tr>
<tr>
<td>Mondelli et al. 2006</td>
<td>Ulnar Neuropathy at Elbow; Carpal Tunnel Syndrome</td>
<td>n= 292</td>
<td>Ulnar neuropathy at the elbow Questionnaire</td>
<td>Reliability- Internal consistency &amp; test-retest reliability;</td>
</tr>
<tr>
<td>Murphy et al. 2011</td>
<td>Charcot-Marie-Tooth disease</td>
<td>n= 34</td>
<td>Charcot-Marie-Tooth disease neuropathy score- 2</td>
<td>Reliability- inter-rater &amp; intra-rater reliability</td>
</tr>
<tr>
<td>Novak et al. 2010</td>
<td>Peripheral Nerve injury</td>
<td>n= 124</td>
<td>The Disabilities of Arm, Shoulder and Hand Questionnaire</td>
<td>Reliability- Internal consistency;</td>
</tr>
<tr>
<td>Novak et al. 2004</td>
<td>Type 2 diabetic NeP</td>
<td>n= 30</td>
<td>Foot Function Index (pain sub scale)</td>
<td>Reliability- Internal consistency;</td>
</tr>
<tr>
<td>Oerlemans et al. 2000</td>
<td>Reflex Sympathetic Dystrophy Syndrome</td>
<td>n= 54</td>
<td>The Radboud skills Questionnaire</td>
<td>Reliability- inter-rater &amp; test-retest reliability;</td>
</tr>
<tr>
<td>Padua et al. 2008</td>
<td>Charcot-Marie-Tooth disease</td>
<td>n= 211</td>
<td>Barthel Index; Deambulation Index</td>
<td>Validity- Construct validity</td>
</tr>
<tr>
<td>Study</td>
<td>Condition</td>
<td>Subjects</td>
<td>Gender (M:F)</td>
<td>Measures</td>
</tr>
<tr>
<td>---------------</td>
<td>------------------------------------------------</td>
<td>----------</td>
<td>--------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Padua et al. 2009</td>
<td>Peripheral Nerve disease</td>
<td>n=392</td>
<td>218:174</td>
<td>Italian Neuropathic Pain Symptom Inventory</td>
</tr>
<tr>
<td>Perez et al. 2002</td>
<td>Complex regional pain syndrome-1</td>
<td>n=21</td>
<td>4:17</td>
<td>Walking stairs Questionnaire; Questionnaire rising and sitting down</td>
</tr>
<tr>
<td>Rejas et al. 2008</td>
<td>Neuropathic Pain</td>
<td>n=603</td>
<td>211:392</td>
<td>Sheehan Disability Scale</td>
</tr>
<tr>
<td>Schmader et al. 2007</td>
<td>Herpes Zoster</td>
<td>n=165</td>
<td>66:99</td>
<td>Zoster Impact Questionnaire; Zoster Brief Pain Inventory</td>
</tr>
<tr>
<td>Sezgin et al. 2006</td>
<td>Idiopathic Carpal Tunnel Syndrome</td>
<td>n=67</td>
<td>5:62</td>
<td>Turkish version of the Boston Questionnaire</td>
</tr>
<tr>
<td>Shy et al. 2005</td>
<td>Charcot-Marie-Tooth disease</td>
<td>n=60</td>
<td>not mentioned</td>
<td>Charcot-Marie-Tooth disease neuropathy score</td>
</tr>
<tr>
<td>Shy et al. 2008</td>
<td>Charcot-Marie-Tooth disease</td>
<td>n=72</td>
<td></td>
<td>Charcot-Marie-Tooth disease</td>
</tr>
</tbody>
</table>
For Peer Review

Singleton et al. 2008 Diabetic peripheral NeP n= 129 Gender = not mentioned Neuropathy Score; Neuropathy Impairment Score Reliability- inter-rater reliability; Validity- Criterion validity, Responsiveness

Smith et al. 2004 Diabetic peripheral NeP n= 57 Gender = 57 M, 0 F Step Activity Monitor Validity- Hypothesis testing

Solari et al. 2008 Charcot-Marie-Tooth disease n= 40 Gender = 21 M, 19 F The Overall Neuropathy Limitations Scale; 10 m walk; 9 hole peg test Reliability- inter-rater & intra-rater reliability

The SALSA Group Leprosy & Diabetes related NeP n= 568 Screening of Activity Limitation and Safety Awareness Scale Reliability- Internal consistency; Validity- Content validity

Valk et al. 2000 Type I and II Diabetes NeP n= 78 The Diabetes symptom checklist- Type 2 Reliability- test-retest reliability; Validity- Construct validity

van Schie et al. 2011 Diabetic peripheral neuropathy n= 24 Step Activity Monitor (4 minute walking test) Validity- Construct validity & Criterion validity

VanSwearingen & Facial paralysis n= 46 Facial Disability Index Reliability- Internal Consistency;
<table>
<thead>
<tr>
<th>Study</th>
<th>Condition</th>
<th>Sample Size</th>
<th>Instrument(s)</th>
<th>Reliability/Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brach 1996</td>
<td>Gender: 16 M, 30 F</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Videler et al. 2008</td>
<td>Hereditary motor and sensory type 1a neuropathy</td>
<td>n = 49</td>
<td>Sollerman Hand function test; Functional dexterity test</td>
<td>Reliability- Internal Consistency &amp; test-retest</td>
</tr>
<tr>
<td>Villoria et al. 2011</td>
<td>Chronic Neuropathic Pain</td>
<td>n = 548</td>
<td>Spanish Neuropathic Pain Symptom Inventory</td>
<td>Reliability- Internal Consistency, test-retest</td>
</tr>
<tr>
<td>Zelman et al. 2005</td>
<td>Diabetic Peripheral NeuP</td>
<td>n = 255</td>
<td>Brief Pain Inventory- Diabetic Peripheral Neuropathy scale</td>
<td>Reliability- Internal Consistency; Validity- Construct validity, Discriminative &amp; Criterion validity</td>
</tr>
<tr>
<td>Zimmerman et al. 2009</td>
<td>Ulnar nerve injury</td>
<td>n = 48</td>
<td>The Disabilities of the Arm Shoulder and Hand Questionnaire; Levine-Katz Questionnaire</td>
<td>Validity- Criterion validity &amp; Construct validity</td>
</tr>
</tbody>
</table>

Pain Practice
**Table III** The COSMIN checklist with 4-point scale [Terwee 2012]

<table>
<thead>
<tr>
<th>Step 1</th>
<th>Evaluated measurement properties in the article: Internal consistency, Reliability; relative measures (including test-retest reliability, inter-rater reliability and intra-rater reliability), Measurement error; absolute measures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Content validity (including face validity), Structural validity, Hypothesis testing, Cross-cultural validity, Criterion validity, Responsiveness and Interpretability</td>
</tr>
<tr>
<td>Step 2</td>
<td>Determining if the statistical method used in the article are based on Classical Test Theory (CTT) or Item Response Theory (IRT): Box General requirements for studies that applied IRT models: excellent/ good/ fair/ poor</td>
</tr>
<tr>
<td>Step 3</td>
<td>Determining if a study meets the standards for good methodological quality: excellent/ good/ fair/ poor</td>
</tr>
<tr>
<td>Step 4</td>
<td>Determining the Generalizability of the results</td>
</tr>
</tbody>
</table>
### Table IV Summary of identified pain Intensity outcome measures with their published psychometric properties and COSMIN grading

<table>
<thead>
<tr>
<th>OMs</th>
<th>Reliability</th>
<th>COSMIN</th>
<th>Validity:</th>
<th>COSMIN</th>
<th>Responsiveness</th>
<th>COSMIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPI-DPN</td>
<td>Internal consistency: excellent</td>
<td>Construct validity:</td>
<td>fair</td>
<td>xx</td>
<td>xx</td>
<td></td>
</tr>
<tr>
<td>Zelman (2005): BPI-DPN showed satisfactory unidimensionality both for the severity and the interference scales (Excellent, α= 0.94)</td>
<td>Zelman (2005): BPI-DPN showed satisfactory construct validity for both the severity and the interference scales</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DPN</td>
<td>Discriminant validity: excellent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zelman (2005): Subcomponents of BPI-DPN: the severity and the interference scale showed satisfactory discriminant validity as both are correlated to a different extent with other measures- SF-12, and HADS (p&lt; 0.001)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Criterion validity: poor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zelman (2005): BPI-DPN severity scale showed high and significant correlations with SF-12v2, and VRS, r's&gt; 0.66 at p&lt; 0.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Concurrent validity: 
Harden (2010): Higher CRPS scores were significantly associated with higher Rand 36 scores (pain intensity, worse physical and social functioning, greater role limitations due to physical and emotional problems, and lower energy and emotional well-being)

Construct validity: 
Valk (2000): dSCT2 showed appropriate correlation with almost all nerve function tests

Hypothesis testing: 
Novak (2004): FFI pain subscale showed moderate correlation with 6 meter walk test ($r = -0.449$, $p<0.001$)

Construct validity: 
Padua (2009): I-NPSI scores showed significant responsiveness
agreement between I-NPSI scores at two different visits correlate with DN4, VAS and ID pain changes (p=0.001) represent reliable measurements to assess NeP symptoms and effectiveness of treatment on them

MPQ xx xx

**Concurrent validity:** poor xx xx

Helme (1989): MPQ showed a significant correlation with VAS (r= 0.67), Word descriptor scale (r= 0.67), and ADL measures (r= 0.53, p< 0.001)

mTCNS

**Internal consistency:** good

Bril (2009): mTCNS showed satisfactory unidimensionality (Moderate, α= 0.78)

**Criterion validity:** excellent xx xx

Bril (2009): Low but acceptable correlation with TCNS (Poor, γ= 0.58)

**Inter-rater reliability:** good

Bril (2009): Satisfactory ICC scores with good reliability (ICC= 0.83, 95% CI)

**Intra-rater reliability:** good

Bril (2009): Satisfactory correlation with
symptom and sensory test (κ = 0.55–0.73)

**Hypothesis testing: Descriptive validity:** poor

*Galer (1997):* 10 NPS pain descriptors showed minimal overlap between most items (γ < 0.50)

**Responsiveness:** excellent

*Jensen (2005):* NPS was significantly able to detect changes from pre-treatment to post-treatment scores

**Predictive validity:** poor

*Galer (1997):* From 10 NPS pain descriptors, only four of descriptors (sharp, cold, sensitive and itchy pain) were able to discriminate PHN pain from other sources of pain, α = 0.01 level

*Jensen (2006):* From 10 NPS pain descriptors, seven descriptors (intense, sharp, hot, dull, sensitive, unpleasant, and deep pain) were significantly able to pick up changes in score after treatment

**Face validity:** fair

**Responsiveness:** poor
Bouhassira (2004): Satisfactory ICC scores with excellent test retest reliability (ICC > 0.90)

Bouhassira (2004): The NPSI was completed accurately and appeared to be fully understood, notably by elderly subjects

Bouhassira (2004): Poor but acceptable correlations with PGIC and CGIC scores

Content validity: excellent

Crawford (2008): Majority of subjects did not raise any concerns with NPSI. Thus no changes to NPSI were consistently suggested

Structural validity: fair

Bouhassira (2004): Each of five factors of NPSI corresponded to a relevant clinical component of NeP

Convergent validity: fair

Bouhassira (2004): Poor but low correlation with global pain intensity measured by a numerical scale (p= 0.60, p < 0.001)

Divergent validity: fair

Bouhassira (2004): No correlation with anxiety and depression scores measured by HADS (p= 0.27; and p=...
Criterion validity: fair

Bouhassira (2004): Lower but acceptable correlations:
pain with brushing ($\rho = 0.70$), pain due to pressure ($\rho = 0.73$); and pain due to cold ($\rho = 0.66$)

Responsiveness: fair

Farrar (2010): On ROC analysis a raw change of -1.74 and a % change of -27.9% were associated with clinically meaningful change

NTSS-6

Internal consistency: poor

Bastyr (2005): NTSS-6 showed satisfactory unidimensionality (Moderate, $\alpha = 0.7$)

Construct validity: fair

Bastyr (2005): NTSS-6 and NSC scores showed moderately positive and significant correlation. ($\Upsilon = 0.773-0.885$, $p < 0.001$)

Responsiveness: MCIDs fair

Bastyr (2005): A change of 0.97 points showed a reasonable change for
<table>
<thead>
<tr>
<th>Component</th>
<th>Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test-retest reliability:</strong></td>
<td><em>Fair</em></td>
</tr>
<tr>
<td><em>Bastyr (2005):</em> Satisfactory ICC scores with lower but acceptable test retest reliability (Baseline ICC= 0.900, End point ICC= 0.903)</td>
<td></td>
</tr>
<tr>
<td><strong>Convergent validity:</strong></td>
<td><em>Fair</em></td>
</tr>
<tr>
<td><em>Bastyr (2005):</em> NTSS-6 and NSC scores showed poorly positive and significant correlation with changes from baseline (γ= 0.519-0.708, p&lt; 0.001)</td>
<td></td>
</tr>
<tr>
<td><strong>Responsiveness:</strong></td>
<td><em>Good</em></td>
</tr>
<tr>
<td><em>Farrar (2001):</em> On ROC analysis a raw change of -2, -2.5, and -3 were associated with least, average, and worst pains</td>
<td></td>
</tr>
<tr>
<td><strong>Internal consistency:</strong></td>
<td><em>Fair</em></td>
</tr>
<tr>
<td><em>Jensen (2010):</em> PQAS showed satisfactory unidimensionality: Deep scale (Moderate α= 0.75), surface scale (Poor α= 0.69), and paroxysmal scale (Good α= 0.87)</td>
<td></td>
</tr>
<tr>
<td><strong>Construct validity:</strong></td>
<td><em>Fair</em></td>
</tr>
<tr>
<td><em>Jensen (2010):</em> Three of the PQAS items and scale scores showed significant correlation with concurrent pain interference on BPI (p&lt; .01)</td>
<td></td>
</tr>
<tr>
<td><strong>Responsiveness:</strong></td>
<td><em>Poor</em></td>
</tr>
<tr>
<td><em>Jensen (2006):</em> Ten of the PQAS descriptor items significantly picked up the changes in scores after treatment (p&lt; .0025)</td>
<td></td>
</tr>
</tbody>
</table>
**P-NPSI**

*Test-retest reliability:* fair

*Face validity:* poor

*Responsiveness:* fair

*Construct validity:* poor

---

**de Andrade (2011):** Satisfactory ICC scores with moderate test retest reliability (ICC = 0.7678)

---

**Face validity:**

*de Andrade (2011):* P-NPSI was filled in less than 8 minutes by 85% of participants. Prevalence rate = 65%

**Responsiveness:**

*de Andrade (2011):* PV-NPSI change scores show significant correlation with P-GIC (Good $\rho = 0.727$), and C-GIC scores (Poor $\rho = 0.645$)

**Construct validity:**

*de Andrade (2011):* PV-NSSI showed low but acceptable correlation with NRS: at first visit (Poor $\rho = 0.40$, $p < 0.0001$), at second visit (Poor $\rho = 0.53$, $p < 0.0001$), and change score (Poor $\rho = 0.22$, $p < 0.0001$)

---

**QST**

*Inter-rater reliability:* good

*Construct validity:* poor

---

**Geber (2011):** QST showed significant inter-rater reliability, $r = 0.83$ (range = 0.56 - 0.89, $p < 0.01$)

**Felix (2009):** QST showed significant correlation with average thermal pain threshold ($r = 0.58$ at $p < 0.02$)

---

**Maser (1989):** 81% of inter-observer agreement that QST can be used adjacent to clinical examination for NeP assessment
**test-retest reliability:** poor

*Felix (2009):* Low but acceptable ICC scores: cold, and hot pain (Poor ICCs = 0.50)

*Geber (2011):* QST showed significant test-retest reliability, $r = 0.86$ (range = 0.67 - 0.93, $p < 0.01$)

**Construct validity:**

*Schreuders (2008):* SESWM showed low but significant correlations with MMT (Poor $r = 0.57$), RIHM dynamometry (Poor $r = 0.70$), and dexterity (Poor $r = 0.65$, $p < 0.001$)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Internal consistency</th>
<th>Construct validity</th>
<th>Responsiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF-MPQ-2</td>
<td>excellent</td>
<td>fair</td>
<td>fair</td>
</tr>
</tbody>
</table>

*Dworkin (2009):* SF-MPQ-2 showed satisfactory unidimensionality: Web survey data (Excellent, $\alpha = 0.91$), and clinical trial data (Excellent, $\alpha = 0.95$)  

*Dworkin (2009):* SF-MPQ-2 scores showed significant correlation with rating of pain and sleep interference, BPI interference scale scores, the SF-36 PCS, MCS scores, the HADS anxiety and depression subscale  

*Dworkin (2009):* Both total and sub-scale scores were responsive to changes that were meaningful to
<table>
<thead>
<tr>
<th>Spanish</th>
<th><strong>Internal consistency:</strong></th>
<th><strong>Construct validity:</strong></th>
<th><strong>test-retest reliability:</strong></th>
<th><strong>Inter-rater reliability:</strong></th>
<th><strong>Construct validity:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>NPSI</td>
<td>S-NPSI showed satisfactory unidimensionality: total NPSI score ($\alpha &gt; 0.80$), and NPSI sub scores ($\alpha &gt; 0.70$)</td>
<td><em>Villoria (2011):</em> S-NPSI showed acceptable accuracy to detect responses of pain as defined by either the clinical or the discriminant criteria</td>
<td><em>Villoria (2011):</em> Moderate test-retest reliability with satisfactory ICC scores (0.680-0.810)</td>
<td><em>Bril (2002):</em> Low but acceptable inter-rater reliability (6.3%)</td>
<td><em>Bril (2002):</em> TCSS showed poor and inverse correlation with SUMAMP and SUMCV ($\Upsilon = 0.424$; $\Upsilon = 0.302$ at $p&lt; 0.0001$; and $p=0.0044$)</td>
</tr>
<tr>
<td>TCSS</td>
<td><em>Bril (2002):</em> Moderate and satisfactory intra-rater reliability (7.3%)</td>
<td><em>Bril (2002):</em> TCSS showed poor and inverse correlation with SUMAMP and SUMCV ($\Upsilon = 0.424$; $\Upsilon = 0.302$ at $p&lt; 0.0001$; and $p=0.0044$)</td>
<td><em>Bril (2002):</em> Moderate and satisfactory intra-rater reliability (7.3%)</td>
<td><em>Bril (2002):</em> Moderate and satisfactory intra-rater reliability (7.3%)</td>
<td><em>Bril (2002):</em> Moderate and satisfactory intra-rater reliability (7.3%)</td>
</tr>
</tbody>
</table>
**TNS**

**Inter-rater reliability:** fair

*Cornblath (1999):* Satisfactory ICC scores with excellent inter-rater reliability (ICC= 0.938, 95% CIs, p≥ 0.836)

**Intra-rater reliability:** fair

*Cornblath (1999):* Satisfactory ICC scores with excellent intra-rater reliability (ICC= 0.973, 95% CIs, p≥ 0.950)

**Construct validity:**

*Cornblath (1999):* TNS showed significantly high and positive correlation with NIS (Good, r= 0.89, 95% CIs) & NSS (Good, r= 0.86, 95% CIs)

**TRNDSI**

**Test-retest reliability:** good

*Collins (2008):* Satisfactory test-retest reliability for CRPS-I and Fibromyalgia (Excellent and Good, ICC= 0.93; and 0.83)

**Measurement error:** good

*Collins (2008):* SEM values were small
compared with domain sum scores (3.5%- 8.3%)

**UENS inter-rater reliability:** poor
Singleton (2008): UENS showed a satisfactory high inter-rater reliability (94%)

**Criterion validity:** excellent
Singleton (2008): UENS (baseline and changeover scores) showed a close correlation with Michigan Diabetic Neuropathic scale and Neuropathy Impairment Score- Lower Leg (p< 0.001)

**Responsiveness:** poor
Singleton (2008): UENS showed a Good diagnostic sensitivity at baseline without sacrificing specificity

**Hypothesis testing:**
Davidoff (1988): The VAS had significant correlations with limb volume ($r^2 = 0.160$), active ROM (upper extremity: $r^2 = 0.167$; lower extremity: $r^2 = 0.508$) and joint pain ($r^2 = 0.341$)

**ZBPI test-retest reliability:** good
Coplan (2004): ZBPI showed low but acceptable test-retest reliability (Poor, ICC= 0.63 b/w 5-7 days; Moderate, ICC=)

**Hypothesis testing:**
good
Coplan (2004): ZBPI showed satisfactory and acceptable correlations with MPQ (24 hours: $\gamma > 0.79$ and for 14-35 days $\gamma > 0.65$), ADL (for 14-35 days: $\gamma$
0.78 b/w 8-10 days and 11-14 days after rash onset) >0.52), and QoL (γ= 0.78)

Schmader (2007): ZBPI showed a significant correlation with other domains. Increased composite pain and discomfort intensity scores were associated with increase in ZBPI ADL interference

Abbreviations: ADL= Activities of Daily Living, BPI= Brief Pain Inventory Scale, BPI- DPN= Brief Pain Inventory Scale for Diabetic Peripheral Neuropathy, CGIC= Clinical Global Impression of Change, CPD= Chronic pain descriptors, CRPS= Complex Regional Pain Syndrome severity, DN4= Douleur Neuropathique 4, dSCT-2= diabetes Symptom Checklist Type-2, FFI= Foot Function Index, HADS= Hospital Anxiety and Depression Scale, QoL= Quality of Life, LANSS= Leeds Assessment of Neuropathic pain Symptoms and signs Screening Tool, MCS= Mental Component Summary, MPQ= McGill Pain Questionnaire, mTCNS= modified Toronto Clinical Neuropathy Score, NIS= Neuropathy Impairment Score, NPS= The Neuropathic Pain Scale, NPSI= Neuropathic Pain Sensory Inventory, NRS= Numeric Rating Scale, NSC= Neuropathy Symptom and Change score, NSS= neuropathy sensory symptoms, NTSS-6= Neuropathy Total Symptom Score-6, PGIC= Patient Global Impression of Change, PI-NRS= Pain Intensity Numeric Rating Scale, P-NPSI= Portuguese version of the Neuropathic Pain Symptoms Inventory, QST= Quantitative Sensory Testing, RIHM= Rotterdam Intrinsic Hand Myometer, SEM= Standard Error of Mean, SESWM= Sensory evaluation with Semmens-Weinstein Monofilaments, SF-MPQ= Short-form McGill Pain Questionnaire, SF-12= The Medical Outcomes Study Short Form Health Survey (SF-12), SUMAMP= Sum of lower limb distal amplitude, TCSS= Toronto clinical scoring system, TNS= Total Neuropathy Score, TRNDSI= The Trauma Related Neuronal Dysfunction Symptoms Inventory, UENS= The Utah Early Neuropathy Scale, VAS= Visual Analog Scale, VRS= Verbal Rating Scale, xx= not determined, ZBPI= Zoster Brief Pain Inventory
<table>
<thead>
<tr>
<th>OMs</th>
<th>Reliability</th>
<th>COSMIN</th>
<th>Validity</th>
<th>COSMIN</th>
<th>Responsiveness</th>
<th>COSMIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMHF</td>
<td>Internal consistency:</td>
<td>poor</td>
<td>Convergent validity:</td>
<td>poor</td>
<td>xx</td>
<td>xx</td>
</tr>
<tr>
<td>Q</td>
<td>Alderson (1999): AMHFQ showed satisfactory unidimensionality (Excellent, α= 0.97)</td>
<td></td>
<td>Alderson (1999): Poor correlation with dynamic two-point discrimination (Y= -0.32), static two-point discrimination (Y= -0.127), the Valpar upper extremity range of motion (Y= -0.2388), Pain VAS (Y= 0.36), functional VAS (Y= 0.3688), grip strength (Y= 0.3867), three point pinch strength (Y= 0.295), and lateral pinch strength (Y= 0.151)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BI</td>
<td>xx</td>
<td>xx</td>
<td>Construct validity:</td>
<td>fair</td>
<td>xx</td>
<td>xx</td>
</tr>
<tr>
<td>Alderson (1999): All the items showed consistent results with in 95th percentile confidence limits (Poor – Moderate ICCs)</td>
<td>Padua (2008): Significant relationship b/w ability to walk on toes, strength of lower limbs muscles, abnormal stand-up, abnormal Romberg test, tactile sensory tests; medium relationship with ability to stand up and strength forearm and intrinsic hand</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table V* Summary of identified physical functioning outcome measures with their published psychometric properties and COSMIN grading
muscles; and lowest relationship with strength of hand intrinsic muscles

**Hypothesis testing:**

*Erdmann (2005):* High BBS showed low correlation with 10 MWT and SIP68 scores (ρ = -0.76, and ρ = -0.62)

**Internal consistency:**

*Lee (2010):* BPI-Facial showed satisfactory unidimensionality: entire instrument (Excellent α = 0.94), intensity of pain (Good α = 0.86), interference with general activities (Good α = 0.89), and interference of facial-specific items (Excellent α = 0.95)

**Construct validity:**

*Lee (2010):* BPI-Facial showed borderline significant correlation with NRS: At least amount of pain (1.01, p = 0.111), and during the week (0.95, p = 0.101)

**Inter-rater reliability:**

*Shy (2005):* Satisfactory ICC scores with excellent inter-rater reliability

**Construct validity:**

*Shy (2005):* CMTNS showed strong and satisfactory correlations with Ambulation Index (r = 0.81), Self-

**Responsiveness:**

*Shy (2008):* CMTNS can be used satisfactorily to detect
(ICC = 0.98, p<0.01) Assesment Questionnaire (r= 0.76), Hand Function progression of CMT disease 
(r= 0.66), 9 Hole Peg test (r= 0.65), CMTNS ulnar and 
median CMAP amplitudes (r= 0.76, 0.72) and 
Neuropathy Impairment Score (r= 0.96)

intra-rater reliability: fair

Shy (2005): The scores from intra-scoring examination did not 
significantly vary on sensory 
evaluation

inter-rater reliability: poor

Murphy (2011): Satisfactory ICC scores 
with excellent inter-rater reliability:
CMTSS2 (ICC= 0.97), and CMTES2
(ICC= 0.96)

intra-rater reliability: poor

Murphy (2011): Satisfactory ICC scores 
with excellent intra-rater reliability:
CMTSS2 (ICC= 0.96), and CMTES2
(ICC= 0.97)
<table>
<thead>
<tr>
<th>DASH</th>
<th><strong>Internal consistency:</strong></th>
<th>poor</th>
<th><strong>Construct validity:</strong></th>
<th>poor</th>
<th>xx</th>
<th>xx</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dias (2008):</td>
<td>DASH showed satisfactory</td>
<td></td>
<td>Dias (2008):</td>
<td>DASH showed no significant correlations</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>unidimensionality (Excellent, $\alpha=0.98$)</td>
<td></td>
<td>with Gartland and Worley scores ($\Upsilon=-0.33$, 5% level)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Novak (2010):</td>
<td>DASH showed</td>
<td>poor</td>
<td>Zimmerman (2009):</td>
<td>DASH showed a significant</td>
<td>fair</td>
<td></td>
</tr>
<tr>
<td></td>
<td>satisfactory unidimensionality</td>
<td></td>
<td>correlation with grip strength ($r=-0.53$), and pinch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Excellent, $\alpha=0.96$)</td>
<td></td>
<td>strength ($r=-0.49$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>test-retest reliability:</strong></td>
<td>poor</td>
<td>Novak (2010):</td>
<td>DASH showed a positive correlation</td>
<td>fair</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dias (2008):</td>
<td>Lower test retest reliability (test-retest differences= -4.7</td>
<td>poor</td>
<td>with VAS for pain (Poor, $r=0.51$, $p&lt;0.001$)</td>
<td>fair</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>to 4.9, 95% CIs, $p=0.02$</td>
<td></td>
<td><strong>Criterion validity:</strong></td>
<td>fair</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Zimmerman (2009):</td>
<td>DASH scores corresponded</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>strongly with clinical staging ($p&lt;0.001$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Hypothesis testing:</strong></td>
<td>poor</td>
<td>Eklund (2009):</td>
<td>DASH showed strong relationship b/w</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>reduced hand function and upper-limb disability:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>manual dexterity ($r=-0.64$), finger dexterity ($r=0.83$),</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>grip strength ($r=-0.72$), tactile gnosis ($r=-0.79$), and</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>hand function index ($r=-0.71$)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>Construct validity:</strong></td>
<td>fair</td>
<td>xx</td>
<td>xx</td>
</tr>
</tbody>
</table>
Padua (2008): DI showed a significant relationship b/w ability to walk on toes, strength of lower limbs muscles, abnormal stand-up, abnormal Romberg test, tactile sensory tests; medium relationship with ability to stand up and strength forearm and intrinsic hand muscles; and lowest relationship with strength of hand intrinsic muscles

<table>
<thead>
<tr>
<th>Test</th>
<th>Reliability</th>
<th>Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMM</td>
<td>fair</td>
<td></td>
</tr>
<tr>
<td>PUT</td>
<td>fair</td>
<td></td>
</tr>
<tr>
<td>FDI</td>
<td>fair</td>
<td></td>
</tr>
<tr>
<td>FDT</td>
<td>fair</td>
<td></td>
</tr>
</tbody>
</table>

*Hypothesis testing: Discriminative validity* - Amirjani (2011): DMMPUT was significantly able to differentiate between impaired hand functions with mild, moderate and severe CTS

*Internal consistency* - VanSwearingen (1996): FDI showed a satisfactory unidimensionality (Theta reliability= 0.88)

*Construct validity* - VanSwearingen (1996): FDI physical function subscale showed a good correlation with clinician’s physical examination of facial movements
with good test retest reliability (ICC= 0.83-0.95, 95% CIs)

HAP

**Internal consistency:** poor

*Farrell (1996):* HAP showed satisfactory unidimensionality (Excellent to Moderate $\alpha= 0.73$- 0.97)

**Hypothesis testing:** fair

*Farrell (1996):* HAP showed strong relationship with both maximum activity score and adjusted activity score (Excellent, $r= 0.97$, $p< 0.000$)

**Responsiveness:** poor

*Farrell (1996):* HAP was sensitive enough to pick up changes in initial scores at the time of discharge

**Criterion validity:** good

*Farrell (1996):* HAP showed strong correlation with maximum activity score (Good $r= 0.78$, $p< .000$), adjusted activity score (Good $r= 0.83$, $p< 0.000$), and Barthel Index: Self-care (Moderate $r= 0.75$, $p< 0.000$), mobilising (Poor $r= 0.61$, $p< 0.000$)

**Concurrent validity:** poor

*Merkies (2006):* INCAT ODSS showed low but significant association with changes in ODSS (Poor $r= 0.66$, $p= 0.007$), Rankin changes (Poor $r=0.60$, $p=0.02$), and GBS Disability Scale changes (Poor $r= 0.56$, $p= 0.04$)
**Internal consistency:** poor

Merkies (2000): ISS showed satisfactory unidimensionality: First visit (Poor $\alpha=0.68$), second visit (Moderate $\alpha=0.73$), third visit (Moderate $\alpha=0.71$), and longitudinal (Good $\alpha=0.87$).

**Construct validity:** fair

Merkies (2000): ISS showed moderate correlations with the additional scales in the stable group (Poor, $r=0.38-0.56$, $p<0.006$).

**Responsiveness:** poor

Merkies (2000): ISS showed significant association of patient's grading with the clinical judgment scores during follow up ($p<0.0001$).

**Inter-rater reliability:** fair

Merkies (2000): Satisfactory ICC scores with good inter-rater reliability (ICC=0.85 to 0.89, $p<0.0001$).

**Intra-rater reliability:** fair

Merkies (2000): Satisfactory ICC scores with good intra-rater reliability (ICC=0.85 to 0.89, $p<0.0001$).

**Criterion validity:** poor

Zimmerman (2009): LKQ showed a significant correlation with DASH: symptom score ($r=0.79$), and
For Peer Review

function score \( r = 0.87, p < 0.001 \)

**Construct validity:** poor

Zimmerman (2009): LKQ function and symptom scores corresponded strongly with clinical staging \( p < 0.001 \)

**Construct validity:** poor

MHQ

**Internal consistency:** poor

Dias (2008): MHQ showed satisfactory unidimensionality (Excellent, \( \alpha = 0.93 \))

**Construct validity:** poor

Dias (2008): MHQ showed no significant correlations with Gartland and Worley scores \( \Upsilon = -0.30, 5\% \text{ level} \)

**test-retest reliability:** poor

Dias (2008): Lower test retest reliability \( \text{test-retest differences} = -4.3 \) to 2.2, 95\% CIs, \( p = 0.02 \)

**Criterion validity:** good

Asad (2010): mNDS proved 92.31\% sensitivity and 47\% specificity in assessing the sensorimotor neuropathy

**Hypothesis testing:** poor

Erdmann (2005): High 10 MWT scores correlated significantly with high SIP68 scores \( p = 0.59, p = 0.036 \)

xx

xx

xx

xx

xx

xx

xx
intra-rater reliability:  fair

Solari (2008): Satisfactory intra-rater reliability with ICC= 0.96 (CI= 0.87-0.99)

inter-rater reliability:  fair

Solari (2008): Satisfactory inter-rater reliability with ICC= 0.95 (CI= 0.89-0.97)

intra-rater reliability:  fair

Solari (2008): Satisfactory intra-rater reliability with ICC= 0.95 (CI= 0.89-0.97)

Responsiveness:  poor

Shy (2008): NIS can be used satisfactorily to detect progression of CMT disease
<table>
<thead>
<tr>
<th>Measure</th>
<th>ODSS</th>
<th>ONLS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inter-rater reliability:</strong></td>
<td><strong>Fair</strong></td>
<td><strong>Poor</strong></td>
</tr>
<tr>
<td>Merkies (2002): Satisfactory ICC scores with excellent inter-rater reliability:</td>
<td>Experienced examiners (ICC= 0.95), Variable examiners (ICC= 0.90)</td>
<td>Graham (2006): Satisfactory ICC scores with excellent test retest reliability (ICC= 0.97)</td>
</tr>
<tr>
<td><strong>Intra-rater reliability:</strong></td>
<td><strong>Fair</strong></td>
<td><strong>Poor</strong></td>
</tr>
<tr>
<td>Merkies (2002): Satisfactory ICC scores with excellent intra-rater reliability:</td>
<td>Experienced examiners (ICC= 0.95), Variable examiners (ICC= 0.93)</td>
<td></td>
</tr>
<tr>
<td><strong>Construct validity:</strong></td>
<td><strong>Fair</strong></td>
<td><strong>Poor</strong></td>
</tr>
<tr>
<td>Merkies (2002): ODSS showed low correlation with MRC (Poor r= 0.45), INCAT sensory sum score (Poor r= 0.41), and Right &amp; left hand grip strengths (Poor r= 0.54 &amp; 0.53)</td>
<td>Graham (2006): ONLS showed a variable correlation with ODSS (Excellent, r= 0.97, p&lt;0.001), 10-meter walk time (Poor, r= 0.58), and MRC score (Poor, r= -0.62)</td>
<td></td>
</tr>
<tr>
<td><strong>Responsiveness:</strong></td>
<td><strong>Good</strong></td>
<td><strong>Poor</strong></td>
</tr>
<tr>
<td>Merkies (2002): Scores showed significant association with clinical changes during follow ups (Poor r= 0.66, p= 0.008)</td>
<td>Graham (2006): ONLS was capable enough to capture a change in activity measures to a similar extent as that of ODSS (SRM= 0.76, 95% CIs)</td>
<td></td>
</tr>
</tbody>
</table>
Solari (2008) Satisfactory inter-rater reliability with weighted kappa for arm score = 0.65 (95% CI = 0.44-0.86), and weighted kappa for leg score = 0.63 (95% CI = 0.41-0.85)

intra-rater reliability:
Solari (2008): Satisfactory intra-rater reliability with weighted kappa for arm score = 0.75 (95% CI = 0.54-0.96), and weighted kappa for leg score = 0.68 (95% CI = 0.47-0.90)

test-retest reliability:
Graham (2006): ONLS showed acceptable test-retest reliability as 15 neurologists independently preferred ONLS

Internal consistency:
Dias (2008): PEM showed satisfactory

Construct validity:
Dias (2008): PEM showed no significant correlations
unidimensionality (Excellent, $\alpha=0.94$) with Gartland and Worley scores ($\Upsilon=-0.37$, 5% level)

**Test-retest reliability:** poor

*Dias (2008):* Lower test retest

reliability (test-retest differences= -9.3 to 2.3, 95% CIs, $p=0.02$)

PPMs

**Test-retest reliability:** poor

*Manor (2008):* Both 6 minute walk

test and Timed up and go test showed

significant reliability (Excellent ICC= 0.93- 0.99, 95% CIs)

QRS

**Test-retest reliability:** poor

*Perez (2002):* QRS showed satisfactory

ICC scores with good test-retest

reliability (range= 0.84- 0.87, $p<0.001$)

RSQ

**Inter-rater reliability:** poor

*Oerlemans (2000):* For inter-rater

reliability the limits of agreement

between two observers was -0.26 and

**Construct validity:** poor

*Oerlemans (2000):* For observer A, 11 test categories

were highly correlated ($>0.80$), however for observer

B, the correlations were lower (but mostly $>0.60$)
test-retest reliability: poor

Oerlemans (2000): For test-retest reliability the limits of agreement between observer A (-0.10 and 0.14) and observer B (-0.26 and 0.22) was very close.

SALSA Internal consistency: excellent

The SALSA Collaborative Study Group (2007): SALSA showed satisfactory unidimensionality: Leprosy group (Good, α= 0.897), and diabetes group (Good, α= 0.814).

Construct validity: poor

Melchior (2011): SALSA showed low but acceptable correlation with NPHT (Moderate r=0.77, p<.0005), SHFE (Poor r= 0.66, p<.0005), and FDT (Poor r= 0.54, p<.005).

Content validity: excellent

The SALSA Collaborative Study Group (2007): SALSA showed strong relationship to the scores assigned by independent experts: Overall (p= 0.67), leprosy group (p= 0.65), and diabetes group (p= 0.70).

Hypothesis testing: poor

SAM xx xx
Smith (2004): SAM showed a strong correlation with physical Function scale, Physical Component Summary score, and Vitality scale (p= 0.01); and a weak correlation with Bodily Pain and Role Limitation (p= 0.05)

Construct validity:
van Schie (2011): SAM (4mWT) showed a significant correlation with Dutch version of International Physical Activity Questionnaire: min/ week (p= 0.49), and activity/ week (p= 0.43, p< 0.05)

Criterion validity:
van Schie (2011): SAM recorded an accuracy of 98.6% compared with observer- counted strides

Internal consistency:
Rejas (2008): SDS showed satisfactory unidimensionality (Excellent, α= 0.904)

Responsiveness:
Rejas (2008): SDS was significantly able to differentiate between responders and non-
**Internal consistency:**

**SHFT**

*Videler (2008)*: SHFT showed excellent homogeneity for both dominant hands ($\alpha=0.96$), and non-dominant hands ($\alpha=0.95$)

**Test-retest reliability:**

*Videler (2008)*: SHFT showed satisfactory test-retest reliability with good ICC (83-0.95, 95% CIs)

**TBQ**

*Sezgin (2006)*: TBQ showed satisfactory unidimensionality: symptom severity scale (Good $\alpha=0.82$), and function status scale (Good $\alpha=0.88$)

**Test-retest reliability:**

*Sezgin (2006)*: TBQ showed satisfactory test-retest reliability with moderate and good correlations with subscales of SF-36: physical functioning ($r=70.55$), physical role ($r=70.54$), bodily pain ($r=70.63$, $p<0.0001$), and emotional role ($r=70.40$, $p<0.001$)

**Construct validity:**

*Sezgin (2006)*: TBQ showed satisfactory correlations with symptoms severity scale ($r=0.73$, $p<0.00001$); moderate and good correlations with subscales of SF-36: physical functioning ($r=70.55$), physical role ($r=70.54$), bodily pain ($r=70.63$, $p<0.0001$), and emotional role ($r=70.40$, $p<0.001$)
satisfactory correlation scores with acceptable test-retest reliability:
symptom severity scale (Poor, $r=0.60$), and function status scale
(Moderate $r=0.77$, $p=0.0001$)

**UNEQ Internal consistency:**
- Poor

  **Content validity:**
  - Mondelli (2006): UNEQ showed satisfactory unidimensionality (Good, $\alpha=0.87$)

  **Responsiveness:**
  - Poor

  **Construct validity:**
  - Mondelli (2006): UNEQ showed significant responsiveness in picking up difference in scores at follow ups (Good, $r=0.85$, $p<0.001$)

  **Hypothesis testing:**
  - Poor

**Walk-12**
- Poor

  **Internal consistency:**
  - Poor

  **Hypothesis testing:**
  - Poor

**Pain Practice**
**test-retest reliability:**

- **poor** Component Summary Score ($r = 20.72$) and the lower limb section of the ONLS ($r = 0.77$)

**Graham (2006):** Satisfactory ICC scores with excellent test retest reliability (ICC= 0.96)

**WSQ test-retest reliability:**

- **poor** $xx$

**Perez (2002):** WSQ showed satisfactory ICC scores with moderate test-retest reliability (range= 0.78-0.87, $p < 0.001$)

**WST test-retest reliability:**

- **poor** $xx$

**Kilmer (2000):** WST showed acceptable test-retest reliability: Pronation (Good ICC= 0.88), supination (Good ICC= 0.85), push (Excellent ICC= 0.96), pull (Excellent ICC= 0.93), and lever arm push (Poor ICC= 0.67)

**Construct validity:**

**poor** $xx$

**Kilmer (2000):** WST showed strong and positive correlations with Hand Held Dynamometry-measured peak torque for both dominant and non-dominant hands ($p < 0.05$)
Hypothesis testing: fair

Schmader (2007): ZIQ showed a significant correlation with other domains. Increased composite pain and discomfort intensity scores were associated with increase in ZIQ ADL interference scores.

Abbreviations: ADL= Activities of Daily Living, AMHFQ= The Alderson-McGall hand function questionnaire, BBM/S= Berg Balance Measure/Score, BI= Barthel Index, BPI= Brief Pain Inventory, CMAP= compound muscle action potential, CMTNS= Charcot-Marie-Tooth disease Neuropathy score, DASH= The Disabilities of Arm, Shoulder and Hand Questionnaire, DI= Deambulation Index, DMMPUT= Dellon-modified Moberg pick-up test, FDI= Facial disability Index, FDT= Functional dexterity test, GBS= Guillain Barré Syndrome, HADS= Hospital Anxiety and Depression Scale, HAP= Human Activity Profile, ISS= Inflammatory neuropathy Sensory Score, LKQ= Levine-Katz Questionnaire, MHQ= The Michigan Hand Outcome Questionnaire, 10-MWT= 10-Meter walking test, mNDS= modified Neuropathy Disability Score, NHPT= Nine-Hole Peg test, NIS= Neuropathy Impairment Score, NRS= Numeric Rating Scale, ODSS= The Overall Disability Sum Score, ONLS= The Overall Neuropathy Limitations Scale, PEM= The Patient Evaluation Measure, PPMs= Physical Performance Measures (6 minute walk test, Timed up and go test), QRS= Questionnaire rising and sitting down, R36HS= Rand-36 Health Survey, RSQ= The Radboud skills Questionnaire, SALSA= Screening of Activity Limitation and Safety Awareness Scale, SAM= Step Activity Monitor, 4mWT= 4 min walk test, SDS= Sheehan Disability Scale, SHFT= Sollerman Hand function test, SIP68= Sickness impact profile 68, TBQ= Turkish version of the Boston Questionnaire, UNEQ= Ulnar neuropathy at the elbow Questionnaire, VAS= Visual Analog Scale, Walk-12= 12-Item Multiple Sclerosis Walking Scale, WSQ= Walking stairs Questionnaire, WST= Work stimulation tasks (knob turn, Linear motion, and Lever arm), xx= not determined, ZIQ= Zoster Impact Questionnaire
### Table VI: Definition of domains, measurement properties, aspects of measurement properties and accepted statistical analyses by COSMIN

<table>
<thead>
<tr>
<th>Domain</th>
<th>Measurement property</th>
<th>Aspect of a measurement property</th>
<th>Definition</th>
<th>Accepted statistical analyses</th>
<th>Interpretation</th>
<th>Inappropriate statistical analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Internal consistency</td>
<td>The degree of the interrelatedness among the items</td>
<td></td>
<td>Cronbach’s alpha (α)</td>
<td>α&gt; 0.90: Excellent</td>
<td>Pearson’s correlation coefficient</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Internal consistency coefficient</td>
<td>α= 0.81- 0.90: Good</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>α&gt; 0.71-0.80: Moderate</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>α&lt; 0.70: Poor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reliability</td>
<td></td>
<td></td>
<td></td>
<td>Continuous scores: ICC</td>
<td>ICC or K&gt; 0.90: Excellent</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Dichotomous/nominal scores: Cohen’s kappa (K)</td>
<td>ICC or K=0.81-0.90: Good</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ordinal scores:</td>
<td>ICC or K&gt; 0.71-0.80:</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Weighted kappa</td>
<td>Moderate</td>
<td>ICC or K&lt; 0.70: Poor</td>
</tr>
<tr>
<td>Reliability</td>
<td></td>
<td></td>
<td></td>
<td>SEM, SDC or LoA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content validity</td>
<td></td>
<td></td>
<td></td>
<td>The degree to which the content of a HR-PRO is an adequate reflection of the construct to be measured</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content validity</td>
<td></td>
<td></td>
<td></td>
<td>Requires a subjective</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content validity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
of) an instrument indeed looks as though they are an adequate reflection of the construct to be measured, thus no analytical standards are developed.

**Construct validity**

The degree to which the scores of a HR-PRO are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the HR-PRO instrument validly measures the construct to be measured.

**Structural validity**

The degree to which the scores of a HR-PRO are an adequate reflection of the dimensionality of the construct to be measured.

<table>
<thead>
<tr>
<th>Hypotheses</th>
<th>Idem construct validity</th>
<th>Correlation coefficient</th>
<th>Positive correlation:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>testing-</strong></td>
<td></td>
<td>Correlation coefficient</td>
<td>γ&gt; 0.90: Excellent</td>
</tr>
<tr>
<td><strong>Discriminant</strong></td>
<td></td>
<td>γ= 0.81-0.90: Good</td>
<td></td>
</tr>
<tr>
<td><strong>validity</strong></td>
<td></td>
<td>γ&gt; 0.71-0.80:</td>
<td></td>
</tr>
<tr>
<td><strong>Convergent</strong></td>
<td></td>
<td>Moderate</td>
<td></td>
</tr>
<tr>
<td><strong>validity</strong></td>
<td></td>
<td>γ&lt; 0.70: Poor</td>
<td></td>
</tr>
<tr>
<td><strong>Divergent</strong></td>
<td></td>
<td>Inverse correlation:</td>
<td></td>
</tr>
</tbody>
</table>
validity;
Sensitivity & specificity

$\gamma < -0.90$: Excellent
$\gamma = -0.81$ to $-0.90$: Good
$\gamma = -0.71$ to $-0.80$: Moderate
$\gamma > -0.70$: Poor

Cross-cultural validity
The degree to which the performance of the items on a translated or culturally adapted HR-PRO instrument are an adequate reflection of the performance of the items of the original version of the HR-PRO instrument

Confirmatory factor analyses
Differential item functioning analyses

Criterion validity
The degree to which the scores of an HR-PRO instrument are an adequate reflection of a ‘gold standard’

When both scores are continuous:
Correlation coefficient

When one is continuous score and other is dichotomous:
Area under the ROC

When both scores are dichotomous:
sensitivity & specificity
<table>
<thead>
<tr>
<th>Responsiveness</th>
<th>Interpretability</th>
</tr>
</thead>
<tbody>
<tr>
<td>The ability of an HR-PRO instrument to detect change over time in the construct to be measured</td>
<td>The degree to which one can assign qualitative meaning—i.e., clinical or commonly understood connotations— to an instrument’s quantitative scores or change in scores</td>
</tr>
</tbody>
</table>

When both scores are continuous: Correlation coefficient

When one is continuous score and other is dichotomous: Area under the ROC

When both scores are dichotomous: sensitivity & specificity

Abbreviations: α= Cronbach’s alpha, HR-PRO= Health related- patient reported outcome, ICC= Intra class correlation coefficient, К= Cohen’s Kappa, LoA= Limits of Agreement, MIC= Minimal important change, MID= Minimal important difference, γ= Correlation coefficient, ROC= Receiver operating curve, SDC= Smallest Detectable Change, SEM= Standard Error of Measurement
Figure I  Flow diagram summarising study selection process
<table>
<thead>
<tr>
<th>Reviewer 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>COMMENT</strong></td>
</tr>
<tr>
<td><strong>Introduction</strong></td>
</tr>
<tr>
<td>I think it is preferable to avoid using too many abbreviations e.g. PMP and OM.</td>
</tr>
<tr>
<td>It would be helpful in the introduction to separate out the two concepts of 1) the need to test the psychometric properties of outcome measures e.g. if reliability has been completed did the results indicate that the test is actually reliable and therefore could be recommended for use;[ in methods would be 2) the methods used to test the psychometric properties (e.g. with COSMIN). The objective gets lost within the final paragraph-can I suggest you rephrase as an aim and move the detail on COSMIN to your methods section.</td>
</tr>
<tr>
<td><strong>Method</strong></td>
</tr>
<tr>
<td>Page 4: Line 41 replace ‘has also been activated’ to ‘was activated’; consider rephrasing this sentence as it is not very clear.</td>
</tr>
<tr>
<td>Check end search date – differs between abstract and methods.</td>
</tr>
<tr>
<td>Please clarify line 56 ‘OMs used in intervention trials…’ with the statement on page 5-eligibility criteria which states that cross sectional clinical trials</td>
</tr>
</tbody>
</table>
(what is this??, can you have a cross sectional intervention trial) and cohort studies (so do you mean an uncontrolled intervention study)?
cross sectional studies and the longitudinal cohort studies.

Page 6: You seem to only describe a method to explore the methodological quality of the individual studies; there is no section on how you made a judgement on ‘the evidence for the psychometric properties’ as indicated in your objective on page 4; and there is no method section to describe how the results will be synthesised (so how can you temper the findings on reliability with the quality of the study-e.g. the study reports that the measure is very reliable but the methodological quality is very low).

Thanks for this comment. We concur with the reviewer’s statement here. The required explanation has been added under the section of data extraction and synthesis. A new table - Table VI has been added explaining the information of the criteria used for synthesizing the results of the study.

Results

Page 8: It would be very helpful if you were able to add some description in the text to summarise the physical function outcomes measures so were they self-report, physical performance, measuring ability e.g. steps versus disability. A similar overview of pain (if possible) would be helpful.

Agreed.

Page 9: It would be important for the reader to know the results of the reliability tests as well as the methodological quality of the study which reported on these results (this would help inform some of the statements in your discussion e.g. page 10, line 53-many OMs seem promising’-on what basis?). So which tests were reliable (need to indicate in your methods how you made that judgement).

Thanks for this suggestion. We concur with reviewer’s concern here.

Table VI has been added to the manuscript, explaining about the judgement criteria used for the studies.

Line 31-35-can you provide evidence to support your statement that ‘these measures have been proven for their PMPs’.

We note the reviewer’s concern here.

Reference has been provided in the text along with Table VI.
Page 9: It would be helpful to describe in a separate section the results for each of COSMIN boxes that you used.

Considering the magnitude of the COSMIN (9 boxes of definitions and explanation for each psychometric property for each outcome measure) and the word limit, explaining about the results of the studies in the form of paragraph seemed to a mere replication of the tables and thus was avoided.

No modifications made.

Discussion

I found the discussion challenging to read as the text of the results did not present the results of the psychometric property under test e.g. if reliability was being tested was many of the tests were reliable and then tempering these findings by only using results from the higher quality studies—you may have done this but it is not explicit to me in your reporting. I think the discussion would become more focused if the methods and results were expanded as I have suggested.

We concur with the reviewer’s statement here. But considering the word count, explaining about the results of the studies in discussion seemed to a mere replication of the tables and thus was avoided. However, the important facts which lead to the results and needs to be highlighted are well explained.

Necessary modifications have been made. The suggestion under the methods and results sections have also been accepted.

Reviewer 2

<table>
<thead>
<tr>
<th>COMMENT</th>
<th>EXPLANATION</th>
<th>MODIFICATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Well done. I have annotated the PDF with some minor grammatical errors; otherwise, the manuscript is well done.</td>
<td>Thanks for your feedback. The potential grammatical mistakes have been made in the sections of Abstract,</td>
<td>Necessary modifications have been made.</td>
</tr>
<tr>
<td>corrected as per your advice.</td>
<td>Introduction, Methodology, Results, and Discussion.</td>
<td></td>
</tr>
</tbody>
</table>