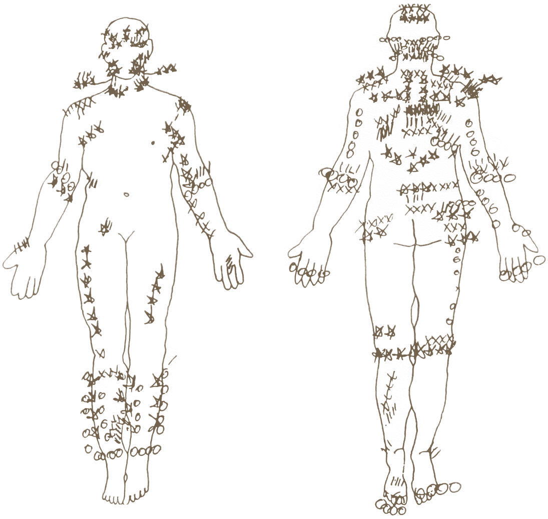
CHAPTER 3

The 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 Jul;66(7):775-82

Ailliet L., Knol D.L., Rubinstein S.M., de Vet H.C.W.,
van Tulder M.W., Terwee C.B.



ABSTRACT

Objectives

To determine the content, structural and construct validity of the Dutch version of the NDI.

Study design and setting

To assess content validity, eleven neck pain experts and ten patients commented on the construct, comprehensiveness and relevance of the NDI. Structural validity was assessed by item factor analysis (FA) and item response theory modelling using the generalized partial credit model. Differential item functioning (DIF) analysis for gender was examined. Pearson's correlation coefficient with the DASH was calculated to assess construct validity.

Results

In addition to a suboptimal translation, we found a lack of consensus on the construct the NDI intends to measure. Experts and patients suggested that the NDI measures more than physical functioning. Unidimensionality of the NDI could not be confirmed. DIF analysis for gender showed DIF for the headache item. The goodness of fit statistics for FA with 1 factor were satisfactory when the item "Concentration" was omitted. A correlation of 0.75 with the DASH was found supporting construct validity.

Conclusions

It is questionable whether in research the NDI should be the instrument of choice for use as a primary outcome measure. Definition of the construct to be measured is a prerequisite for the assessment of validity.

What's new?

1. A clear definition of the construct to be measured is a prerequisite for the assessment of validity. It is unclear what the NDI aims to measure. It measures more than physical functioning, but for the measurement of a broader construct (e.g. disability) important items are missing. This also compromises other aspects of validity.
2. Analysis based on the consensus-based standards for the selection of health measurement instruments (COSMIN Standards) for good content validity shows that the content validity of the NDI is poor. These identified conceptual problems might also apply to other questionnaires developed in the 80's and 90's.
3. In research, the NDI has long been considered the gold standard for measuring disability in patients with neck pain. Given current quality standards for patient reported outcome instruments, questions could be raised whether the NDI should still be considered the first instrument of choice as a primary outcome measure in studies on patients with neck pain. We advocate the development of a new disease-specific instrument, starting from a clear definition of the construct to be measured and using more advanced psychometric techniques.

A recent systematic review on the measurement properties of disease-specific questionnaires in patients with neck pain indicated that the Neck Disability Index (NDI) is the most frequently evaluated questionnaire[1]. Researchers and clinicians often use the NDI to measure the level of disability and to study the effect of an intervention on patients with neck pain[2,3,4,5]. In developing the NDI, Vernon and Mior were inspired by Fairbank and colleagues who developed the Oswestry Disability Index (ODI) for measuring disability in low back pain patients[6]. The NDI was the first instrument designed to assess self-rated disability in patients with neck pain. In non-English speaking countries, researchers and clinicians alike, have to rely on a translated version of the original NDI.

As of late 2007, the NDI had been used in approximately 300 publications and it had been translated into 22 languages[3]. Currently there are over 450 articles internationally that have cited the NDI. Numerous clinical guideline organizations, especially for whiplash management, have endorsed the NDI as the questionnaire of choice for neck pain patients.

Vernon declared in an e-mail conversation (November 2011) that the construct of the NDI was "self-rated disability", where disability was understood as the perceived effect of pain and impairment on the patient's performance and enjoyment of activities of daily living.

However, different opinions exist with regards to the meaning of the construct that the NDI aims to measure. Some researchers interpret the NDI as a measure of functional status[7,8], while others have a broader interpretation and see it as a measure of pain and disability[9]. As a result, confusion might arise as to what the NDI aims to measure and how scores should be interpreted.

The lack of consensus in the construct that the NDI aims to measure might hinder an assessment of the validity of the instrument, as validity is defined as the extent to which an instrument measures what it purports to measure[10].

As far as the Dutch version of the NDI is concerned, it is unclear how and by whom the NDI was translated. There are at least two different Dutch versions of the NDI.

The aim of this article was to evaluate the quality of the translation of the most commonly used Dutch version of the NDI and to determine its content validity, structural validity and construct validity.

METHODS

The Neck Disability Index

The NDI is a patient-reported outcome measure. It consists of 10 items. The 10 items have 6 response categories (range 0-5, total score range 0-50)[11]. (Appendix A)

Translation

In order to check the Dutch translation of the NDI, the most commonly used Dutch version of the NDI[12] was translated back into English by two independently operating professional translators, blinded to the original English version. Discrepancies between the translated version, the versions translated back into English and the original English version were discussed amongst 5 of the 6 authors and with the two professional translators in order to evaluate the quality of the translation of the Dutch version of the NDI.

Population

The NDI was completed by 338 patients with neck pain, who participated in a prospective cohort study with 97 chiropractors in Belgium and the Netherlands. The population included in this study very much resembles the population included in the studies of Hoving[13] and Pool[8]. In those Dutch studies on neck pain in primary care, the patients were recruited by the general practitioner and allocated to usual care, manual therapy, physical therapy and/or behavioral graded activity. Information on the characteristics of the patients included in our study can be found in Table 1. Inclusion criteria were: patients, age 18 to 65 years who had neck pain with or without radiation into the arm as their main complaint, had not consulted a chiropractor for their neck complaint in

TABLE 1: CHARACTERISTICS OF THE 338 PATIENTS, ON WHOSE INFORMATION SUPPLIED IN THE NDI, THIS STUDY WAS BASED

Gender (male / female)	116 / 222 (34.3% / 65.7%)
Age – *Mean [SD]* – *Range*	 41.3 [SD 11.8] 18-65
Total scores NDI – *Mean [SD]* – *Median* – *Range*	 14.9 [SD 6.8] 14 0-32
First episode of neck pain (yes)	17.7%
Duration of the complaint – *< 6 weeks* – *> 6 weeks* – *6 weeks – 3 months* – *> 3 months*	 25.4% 74.6% 15.9% 58.7%
Education – *No high school diploma* – *High school diploma* – *College/university degree* – *Post-university degree*	 25.1% 35.4% 35.6% 3.9%
Referral pattern – *MD / other health care person* – *Family / friends* – *Own initiative*	 25.8% 41.0% 33.2%

NDI = Neck Disability Index, SD = Standard Deviation, MD = Medical Doctor

the past 6 months, and had a good understanding of the Dutch language, both in reading and writing. Exclusion criteria were red flags in the anamnesis or upon clinical examination at the first visit. Patients completed a web-based version of the NDI at baseline – i.e. maximum two days before the initial visit at the chiropractor – and at 1, 3, 6 and 12 months follow-up. Patients were treated for their neck pain by and at the discretion of a chiropractor, but other interventions and/or pain medication were allowed. The 6 months data were used to assess the validity of the Dutch version of the NDI. Part of the patients had finished their chiropractic treatment by then.

Content validity

Based on the consensus-based standards for the selection of health measurement instruments (COSMIN Standards)[14], four requirements for good content validity were defined:

1. All items should refer to relevant aspects of the construct to be measured.
2. All items should be relevant for the study population (e.g. age, gender, disease characteristics, country, setting).
3. All items should be relevant for the purpose of the measurement instrument (discriminative, evaluative, and/or predictive).
4. All items together should comprehensively reflect the construct to be measured.

TABLE 2: CHARACTERISTICS OF THE 10 PATIENTS, RECRUITED TO COMMENT ON THE RELEVANCE OF THE 10 QUESTIONS OF THE NDI

Gender (male / female)	5/5
Age (range)	21-56
Ever treated by a chiropractor before (yes)	3
Mean total score on NDI	21.8
First episode of neck pain (yes)	5
Duration of current episode (<6 weeks/>6 weeks)	4/6

NDI = Neck Disability Index

To evaluate the first requirement, a literature search was performed in Medline (1990 to December 2011), EMbase (1990 to December 2011) and CINAHL (1990 to December 2011) for relevant articles pertaining to the development of the NDI and the definition of disability. Second, contact was established with the developer of the original version of the NDI, H. Vernon, in order to have full clarity as to the exact description of the original construct of the NDI. Third, eleven clinicians and/or researchers with expertise in neck pain, were invited to comment on the construct of the NDI in order to reach a consensus on what the NDI aims to measure. To evaluate the second requirement, ten consecutive new patients presenting to a private chiropractic practice with neck pain as their main complaint were recruited and asked to comment on the relevance of the ten questions of the NDI. The patients were asked to rate the intensity of their current pain on a numeric rating scale (NRS) from 0 to 10. These ten patients were considered representative for the population that participated in the cohort study in which the NDI was used. They were recruited from a representative practice in Belgium (based on a 2010 study on the characteristics of chiropractors and their patients in Belgium[15]). Information on the characteristics of these patients can be found in Table 2. To evaluate the third requirement, the authors then considered if all items were relevant for evaluative purposes, i.e. items should potentially be able to pick up changes when the patient's health status improves or worsens. To evaluate the fourth requirement, both the eleven experts mentioned above and the selected ten patients were invited to comment on the comprehensiveness of the ten questions from the NDI, i.e. whether items were missing.

Structural validity

Prior to assessing structural validity, score distributions were checked. To assess the dimensionality of the item pool, item factor analysis (FA) for ordered categorical items – i.e. FA on the matrix of polychoric correlations – was performed. Item parameter estimates were obtained using the method of weighted least squares with means and variance adjustment (WLSMV) in Mplus 6.12[16]. We checked for residual (polychoric) correlations between items. Factors with eigenvalues greater than 1 were considered for potential separate scales. Interpretation of multifactor solutions was based on the oblique promax rotated factors. A model was considered to have a good fit when the root mean square error of approximation (RMSEA) was 0.06 or lower, the comparative fit index (CFI) 0.95 or higher and the standardized root mean square residual (SRMR) 0.08 or lower[17].

The fit of each subscale (dimension) was then further investigated by means of item response theory (IRT). One of the most flexible IRT models is the generalized partial credit model (GPCM)[18]. In the GPCM, the probability of a score *j* of an item *i* as a function of the latent disability *θ* of a patient is given by the item response function

$$P_{ij}(\theta) = \frac{exp[\alpha_i(j\theta - \sum_{g=1}^{j}\beta_{ig})]}{1 + \sum_{h=1}^{m_i} exp[\alpha_i(h\theta - \sum_{g=1}^{h}\beta_{ig})]} \qquad j = 0,\ldots,m_i$$

where $m_i + 1$ denotes the number of item categories and β_{ij} and α_i are item parameters. The parameter β_{ij} is a category intersection parameter of item *i*, that is, it is the point in which the probability of responding in category $j - 1$ is equal to the probability of responding in category *j*. Finally, α_i is the discrimination parameter that indicates the extent to which the item response is related to the latent scale. In IRT models, the response probabilities of each subject to the individual items are modelled as a function of the latent disability of that subject. This makes IRT models particularly well-suited for analyzing item fit and differential item functioning (DIF). An item shows DIF if the probability of responding in the different categories of the item varies across groups of patients with the same disability level. Because of known gender differences in neck pain[19], DIF between men and women was assessed by using the Lagrange multiplier (LM) statistic based on the difference between observed and expected item scores in the two subgroups[20]. An iterative approach was followed, whereby an item with the largest significant LM test and a difference > 0.10 was given group-specific item parameters until all remaining DIF tests are no longer significant[21].

After modelling DIF, item fit was investigated, again using LM statistics, now separately for each gender subgroup[22]. All GPCM analyses were carried out using MIRT[23].

Construct validity

Hypothesis testing, aiding in assessing the construct validity, was conducted by comparing the scores on the NDI to the scores on the Disabilities of the Arm, Shoulder and Hand (DASH) questionnaire, an instrument designed as a self-administered measure of symptoms and functional status, with a focus on physical function[24]. The DASH was preferred over other questionnaires pertaining to neck pain because in a recent systematic review on the measurement properties of disease-specific questionnaires in patients with neck pain[1] and in an article by the Bone and Joint Decade 2000-2010 Task Force on neck pain and associated disorders[25], the NDI came out as the questionnaire with the best measurement properties. We therefore chose a fairly well validated questionnaire in a closely related body part: the DASH is one of the most often used instruments to measure physical functioning in those with complaints in the upper extremity[26,27]. The DASH consists of thirty questions pertinent to function of the upper extremity, and an optional four questions on high performance Sport/Music and an optional four work-specific questions. Only the thirty questions pertinent to function of the upper extremity were used, and each question was scored on an ordinal scale from 1-5. A moderate to strong positive correlation (Pearson's correlation coefficient $r > 0.40$) between the DASH and the NDI was expected as an indication of construct validity.

RESULTS

Quality of the Dutch translation

A number of problems were identified with the Dutch translation of the NDI. Firstly, the patient instruction that precedes the original version of the NDI is not included in the Dutch version of the NDI. Secondly, for some items, the Dutch translation (or part of it) was considered inadequate. For example, the Dutch translation of the phrase "… but it causes extra neck pain" in the item "personal care" was back translated into English as "although this increases the pain". Saying that personal care causes extra neck pain can imply that there was no pain to start with, whereas the word "increase" implies that there was pain to start with and that a certain activity makes the pain worse. Another example is the Dutch translation for "recreation". In the Dutch translation, the meaning of the word recreation is much broader than was originally intended, thereby possibly overlapping other items within the questionnaire. It allows for the inclusion of non-physical activities, like reading, which already represent a separate item in the NDI. Thirdly, the formulation of some of the Dutch questions is grammatically incorrect. For example, in some sentences the subject matter was lost in the translation process. Fourthly, the sequence of the questions in the Dutch translated version is different from the sequence in the original NDI. Hains et al reported that changing the sequence of the questions did not affect the score on the NDI[28]. Fifthly, the addition of the qualifier "neck" in the NDI[11] in all places (items 1, 2 and 3) where the sole term "pain" was mentioned in the ODI, was not picked up in the Dutch translation: therefore it could be unclear to the patient that the questionnaire was only concerned with the patient's "neck pain". Finally, the professional translators indicated some problems with the original English version as well. For example, the statement "I have severe headaches which come frequently" refers to both pain intensity and frequency, which may be considered inappropriate since it asks to score two different aspects.

Content validity

Review of the literature and personal communication with the developer of the NDI confirmed that the NDI was based on the concept of disability as described in the – at that time – generally accepted WHO classification of “pain, impairment, disability and handicap”. Vernon described in personal e-mail communication (November 2011) the concept of “disability” as the “perceived effect of pain and impairment on the patient’s performance and enjoyment of activities of daily living”. This was used as the underlying basis for the NDI.

The eleven experts gave varying answers to the question what the NDI aims to measure. All experts report the ability to perform functional activities. In addition, pain or pain intensity, a comprehensive term covering the three items concentration, reading, and sleeping, dysfunction/symptoms, and personal care were mentioned as additional yet separate constructs by at least one and for some constructs by up to nine experts.

Based on the answers from the eleven experts, reflecting their own opinion, we concluded that the NDI measures more than just the ability to perform functional activities or activities of daily living. We then asked the experts to critically assess the concept “burden of disease” as an alternative for the construct “disability”, considering that the NDI measures more than just physical functioning. Most items were considered relevant by the experts for the measurement of “burden of disease”, although there was some doubt with regards to the items “pain intensity” and “driving” as to whether these two items could fit under the construct of “burden of disease”.

Also, experts expressed some concerns with regards to the item “recreation” which might be defined and interpreted differently between different age groups. Finally, the item “work” does not make a distinction between household work, computer work, or other jobs with specific job requirements that might eventually be burdensome to the upper dorsal spine and neck.

Ten patients provided their opinion on the relevance and the comprehensiveness of the NDI. Patients scoring 5 or more on the numeric rating scale (NRS) (n=4) found all ten items relevant. Those six people scoring 4 or less on the NRS experienced their neck complaints much more as bothersome as opposed to painful. They felt that most of the items of the NDI (like “personal care”, “lifting”, “work”, “sleeping”, “driving”, and “recreation”) were not applicable or relevant to them, even in such a way that they expressed their concerns whether or not they could (still) be considered neck pain patients.

The authors decided that all items were relevant for evaluative purposes, confirming that the items of the NDI were suitable to assess change over time.

As far as the comprehensiveness of the ten questions of the NDI is concerned, only those four patients scoring 5 or more on the NRS expressed some remarks. For instance, they missed computer work as specific item. They also mentioned household work as a separate item. The experts suggested that to fully capture "burden of disease" a number of items seemed to be missing: sports, mobility, mental issues, dizziness, radiation, the distinction between household work and professional activities. Some of the experts also indicate that the NDI does not take into account the use of over-the-counter or prescription drugs, nor its potential effect on modulating neck pain. Finally, the experts also noted the absence of an item related to computer work.

Structural validity

Data from 338 patients (66% female, ranging in age from 18 to 65) presenting with neck pain to 97 chiropractors in Belgium and the Netherlands were used for statistical analyses.

Checking score distributions showed that not all item categories were used. In total 3 item categories were left unused, the category corresponding to the highest level of difficulty of the items "personal care", "lifting" and "sleeping".

Exploratory item factor analysis for the entire group showed two factors with eigenvalues greater than 1 (4.11 and 1.27), with 41.1% of the variance explained by the first factor. The two factors cumulatively explained 53.8% of the variation within the questionnaire. After promax rotation, one of the factors was difficult to interpret, since only the items "reading", "headaches" and "concentration" loaded on the second factor (Table 3), and of those three, only "concentration" had a high loading. The rotated factors were interpreted as representing physical and non-physical aspects of pain. The correlation between the two factors was 0.51. The largest residual (polychoric) correlations (> 0.16) were found for the pairs Concentration-Headaches (0.21) and Concentration-Reading (0.20). All other residuals were smaller than 0.16.

The 1-factor model did not fit well: the goodness of fit statistics display values for RMSEA of 0.101, for CFI of 0.933 and for SRMR of 0.075. When the item "concentration"

TABLE 3: PROMAX ROTATED FACTOR LOADINGS OF THE NDI ITEMS

ITEMS	PHYSICAL	NON-PHYSICAL
1. Pain intensity	0.630	-0.036
2. Personal care	0.565	-0.065
3. Lifting	0.678	-0.093
4. Reading	0.360	0.406
5. Headaches	0.152	0.350
6. Concentration	-0.207	1.003
7. Work	0.615	0.168
8. Driving	0.711	0.028
9. Sleeping	0.297	0.201
10. Recreation	0.929	-0.121

NDI = Neck Disability Index / Correlation between physical and non-physical factor (after Promax rotation) = 0.51

TABLE 4: LAGRANGE MULTIPLIER TESTS FOR DIF, BEFORE AND AFTER MODELLING FOR DIF

	INITIAL			FINAL		
ITEM	*LM*	*p-value*	*Absolute Difference*	*LM*	*p-value*	*Absolute Difference*
1. Pain intensity	3.78	0.05	0.06	1.43	0.23	0.04
2. Personal care	1.14	0.28	0.02	2.74	0.10	0.03
3. Lifting	6.79	0.01	0.10	–	–	–
4. Reading	0.64	0.42	0.04	3.10	0.08	0.08
5. Headaches	9.07	0.00	0.21	–	–	–
6. Concentration	1.85	0.17	0.06	1.13	0.29	0.05
7. Work	4.66	0.03	0.06	1.74	0.19	0.04
8. Driving	1.38	0.24	0.04	0.01	0.93	0.00
9. Sleeping	6.17	0.01	0.11	–	–	–
10. Recreation	1.04	0.31	0.02	0.03	0.86	0.00

DIF = differential item functioning, LM = value of Lagrange multiplier test, chi-squared with df = 1

TABLE 5: ITEM PARAMETERS AFTER MODELLING FOR DIF

ITEM	α_i	β_{i1}	β_{i2}	β_{i3}	β_{i4}	β_{i5}
1. Pain intensity	0.959	-2.516	-1.385	1.282	2.556	2.829
2. Personal care	0.908	1.250	2.961	2.337	2.099	
3. Lifting						
men	0.908	-0.455	2.411	-0.181	2.079	
women	0.587	-0.597	1.729	-0.543	1.515	
4. Reading	0.697	0.141	-0.616	0.628	2.633	2.862
5. Headaches						
men	0.310	0.181	0.109	0.347	0.066	1.155
women	0.327	-0.427	-0.305	0.161	0.378	0.906
6. Concentration	0.379	-0.090	0.750	1.346	0.666	2.729
7. Work	1.179	-0.171	0.955	2.642	2.656	2.217
8. Driving	1.277	-1.526	-0.521	1.786	3.711	2.222
9. Sleeping						
men	0.491	-0.240	0.636	1.658	1.617	
women	0.376	-0.519	0.758	0.589	0.857	
10. Recreation	1.842	-3.366	0.914	3.000	3.868	3.662

DIF = differential item functioning, α_i = discrimination parameter, β_{i1}-β_{i5} = category intersection parameters

was removed from the model the fit of the 1-factor model was satisfactory (RMSEA was 0.064, CFI was 0.978 and SRMR was 0.050).

For testing DIF, the initial LM statistics for men and women are presented in Table 4. After allowing the items "headaches", "sleep" and "lifting", respectively to have gender group specific item parameters, no more DIF could be observed (Table 4, final). A clear DIF for the "headaches" item could be observed, where women scored worse than men with comparable levels of disability. Also for “sleep” and for ”lifting”, there was DIF, but less pronounced than for the “headaches” item. Item parameters after modelling DIF are shown in table 5. Since women are chosen as the reference group, the mean for women is 0, with standard deviation 1 (SD). For men the mean is -0.494, and the SD is 0.925. The model allows for this difference. However, assuming the same level of disability, three items showed differing values for men and women (women scored worse).

The GPCM with these three gender-specific parameters fitted the data well, with only “concentration” showing borderline significance (0.05) with absolute difference values of 0.17 for men and 0.10 for women. IRT reliability was 0.78 for men and 0.82 for women.

Construct validity – hypothesis testing

A Pearson correlation coefficient of 0.75 was found between the DASH and NDI.

DISCUSSION

The most commonly used Dutch translation of the NDI was considered to be suboptimal. This consideration can partly be explained by a poor translation, but is potentially a result of a less than optimal formulation of items in the original English version as well. Poor translation implies that scores on the Dutch and the English versions of the NDI might not be comparable. It may also affect content validity if translated items are interpreted differently.

Many researchers interpret the questionnaire as a measure of physical functioning[7,8]. However, our study shows that the NDI measures more than physical functioning. Review of the literature and personal communication with the developer of the NDI demonstrated that the NDI was not designed to measure “physical functioning”. The NDI was based on the concept of disability as “the perceived effect of pain and impairment on the patient’s performance and enjoyment of daily living”. It was thought that this was consistent with the generally accepted WHO classification of “pain, impairment, disability and handicap”. Disability was thus seen as a matter of the way in which each individual with a condition interprets the effect of the condition on their own lives. In personal e-mail communication with Vernon (November 2011) this was termed “self-reported disability”[24]. In 1994, Verbrugge and Jette introduced the term “disablement process”. They defined disability as difficulty doing activities in any domain of life (from hygiene to hobbies, errands to sleep) due to a health or physical problem[29]. Both definitions refer to a construct that is much broader than physical functioning.

However, when we considered a broader construct for assessing the content validity of the NDI we found that important aspects are missing, e.g. sports and computer work. Missing items may in part result from the fact that the NDI was developed in a different era, where for instance personal computer use was much less prevalent. This means that the content of the NDI appears incomplete or in part outdated.

Some of the experts also indicated that the NDI does neither take into account the use of over-the-counter and prescription drugs, nor its potential effect on modulating neck pain. In contrast with the ODI where the item on "pain intensity" was graded in the detractors by phrases related to the use of medication, Vernon deliberately choose not to refer to the use of tablets (medication for pain and/or sleep) "since many subjects might not be taking such medications"[3]. In personal e-mail communication (November 2011), Vernon pointed out that pain intensity should be graded on its own, with terms such as mild, moderate, severe. However, interviewing a representative sample of ten patients revealed that some but not all of the patients suffering from neck pain do take pain medication. This supports the claim that it would have contributed to the correctness of the answers if the introduction to the questionnaire would have included an instruction for the patients how to rate the questions in case they were taking medication. Medication use can have an impact, not only on the scoring of the different items of the NDI, but on the interpretation of the scoring as well.

A final problem related to the content of the NDI is the fact that for three items the highest response category – corresponding to the highest level of disability – was left unused. Previous studies have also observed that response categories designed to measure the highest level of neck pain disability are rarely endorsed by respondents[28,30]. Our sample was representative for patients presenting with neck pain to a chiropractor. This suggests that the extreme categories are not legitimate for neck pain patients in primary (chiropractic) care in Belgium and the Netherlands. Those extreme response categories might well be applicable to patients with neck pain presenting to secondary care.

We conclude that there is a lack of understanding as to what the NDI precisely aims to measure. The results of our study suggest that the NDI measures more than just physical functioning, and that for the measurement of a broader construct important items may be missing. Based upon this information, we conclude that the content validity is poor. In a previous review we rated the content validity as good, because we only scored whether patients were involved in the development, which was the case. However, by applying the four COSMIN criteria for content validity, the lack of a clear definition of the construct led to a poor score on content validity. There is thus a need for improving the content of the NDI or a new instrument needs to be developed.

After modelling for DIF between men and women, all items fitted the GPCM, perhaps with the item "concentration" only marginally fitting. The presence of large residual correlations (> 0.16) for the pairs "Concentration-Headaches" and "Concentration-Reading" illustrates as well that there is possibly a second, non-physical, factor or that "concentration" doesn't fit the 1-factor model. Van der Velde et al. used the Rasch method and came to a unidimensional well fitting model after removing two items ("headaches" and "lifting")[31]. However, the Rasch model is a more restricted model than our GPCM which was the model of choice since the discrimination parameters substantially differed. Actually when all discrimination parameters are equal, the GPCM is identical to the Rasch model. The NDI correlated well with the DASH (r = 0.75), a questionnaire measuring a similar construct. Apparently, the content problems have not much impact on the relation to other measures, or the DASH may have similar problems.

Some limitations of our study should be acknowledged. Only one of the consulted experts had English as his mother language. All other experts considered English as their second language. This might be a source of bias, since it appears to us that it is often unclear to the non-native English speaker/reader what is really meant by "disability". Secondly, the conclusions from this study are based on a Dutch translation of the NDI, applied to a cohort of patients presenting with non-specific neck pain to primary chiropractic care in Belgium and the Netherlands. We acknowledge that our findings might in part be culture-specific, thereby hampering their generalizability. However, the problems relating to the lack of clarity about the construct being measured are mainly language-independent. Thirdly, the data for this study were derived from a cohort study in which patients presenting with neck pain to a chiropractor were followed for 12 months. The data for this study were collected at the 6 months follow-up. Asking the patients to evaluate their status at 6 months allows them to assess the items of the NDI for their relevance at different stages of the complaint. It is plausible that patients had improved as a result of the therapeutic intervention, therefore scoring lower on the NDI. This could provide an explanation why a number of item response categories were left unused.

The problems with the NDI should be interpreted in light of the time in which the questionnaire was developed. The NDI, based on the Oswestry Disability Index (ODI), was developed in 1991 in an attempt to document more than the simpler measures of pain severity, location, or duration that were more commonly used at that time for patients with neck pain. The ODI was, according to its developer Fairbank, deliberately focused on physical activities and not on the psychological consequences of acute or chronic pain[6]. Vernon and Mior, the developers of the NDI, retained five items from the ODI and completed the list, based upon information gathered from informal surveys of patients and a small consulting team of health practitioners, with five more items reported to be importantly affected in neck pain patients[3]. Although the development of the NDI has been very important and has helped clinical practice and research to a great extent over the last two decades, this brief description of the development of the NDI suggests that this instrument would not pass the scrutiny of current rigorous guidelines for instrument development. It should be noted that the issues raised here regarding the NDI probably affect many questionnaires developed in the 80's and 90's, often developed by clinicians for clinicians, in an attempt to come up with a way to quantitatively measure the impact of disease on a patient's life. The DASH for instance, used in our study to assess the construct validity of the NDI, is also vulnerable to the same critical remarks. Recently, more attention is paid to content validity and structural validity. Furthermore, the use of more advanced psychometric techniques like IRT, allow for better measurement instruments to be developed. Newly developed "item banks" may become the measurement instruments of the future. Item banks contain a large collection of questions about a particular construct, which have been carefully calibrated using IRT analysis, and can be administered in total, in part, or computer-adaptive. Rose et al[32] conclude that a 10-item computerized adaptive test based on a preliminary item bank can extend the range of measurement substantially (plus 2-3 standard deviations) to an extent that ceiling and floor effects are very unlikely to occur in clinical applications and that measurement precision is improved over a wide range compared with fixed length questionnaires. An example of a collection of item banks is PROMIS (Patient-Reported Outcomes Measurement Information System), which includes item banks for the construct of e.g. pain and physical functioning[32,33]. A Dutch-Flemish translation of PROMIS is currently being completed. These item banks may be used in the future as an alternative for the NDI, or may serve as a basis for developing a new disease-specific instrument.

CONCLUSION

It is unclear what the NDI aims to measure. It measures more than physical functioning. However, for a broader construct important items are missing. The content validity of the NDI is poor. An important lesson from this study is that a clear definition of the construct to be measured is a prerequisite for the assessment of validity. In line with current recommendations, e.g. from the US Food and Drug Administration[34], we conclude that more attention should be given to the definition of the construct and conceptual model when developing new instruments. Given current quality standards for PRO instruments, questions could be raised whether the NDI should be considered the first instrument of choice as a primary outcome measure in studies on patients with neck pain. We advocate the development of a new neck-specific instrument, starting from a clear definition of the construct to be measured and using more advanced psychometric techniques.

REFERENCES

1. Schellingerhout JM, Verhagen AP, Heymans MW, Koes BW, de Vet HC, Terwee CB. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. Qual Life Res Published online: 07 July 2011.
2. Macdermid JC, Walton DM, Avery SA, Blanchard A, Etruw E, McAlpine C, Goldsmith CH. Measurement properties of the Neck Disability Index: a systematic review. JOSPT May 2009;39(5):400-417.
3. Vernon H. The Neck Disability Index: State of the Art, 1991-2008. J Manipulative Physiol Ther 2008;31:491-502.
4. Cleland JA, Childs JD, Whitman JM. Psychometric properties of the Neck Disability Index and Numeric Pain Rating scale in patients with mechanical neck pain. Arch Phys Med Rehabil 2008;89:69-74.
5. Bronfort G, Evans R, Nelson B, Aker PD, Goldsmith CH, Vernon H. A randomized clinical trial of exercise and spinal manipulation for patients with chronic neck pain. Spine 2001;26:788-797.
6. Fairbank JC, Pynsent PB. The Oswestry Disability Index. Spine 2000 Nov 15;25(22):2940-2952.
7. Riddle DL, Stratford PW. Use of generic versus region-specific functional status measures on patients with cervical spine disorders. Physical Therapy 1998;78(9): 951-963.
8. Pool JJ, Ostelo RW, Knol D, Vlaeyen JW, Bouter LM, de Vet HCW. Is a behavioral graded activity program more effective than manual therapy in patients with sub-acute neck pain? Results from a randomized clinical trial. Spine. 2010;35(10):1017-1024.
9. Sterling M, Jull G, Kenardy J. Physical and psychological factors maintain long-term predictive capacity post-whiplash injury. Pain. 2006;122:102-108.
10. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychological Bulletin 1955;52:281-302.
11. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther 1991;14:409-415.
12. Köke AJA, Heuts PHTG, Vlaeyen JWS, Oostendorp R. 1996 Neck Disability Index. Pijn Kennis Centrum academisch ziekenhuis Maastricht. Meetinstrumenten chronische pijn, Maastricht, Nederland. Pp 52-54.
13. Hoving JL, Koes BW, de Vet HCW, vander Windt DAWM, Assendelft WJJ, van Marmeren H, Devillé WLJM, Pool JJM, Scholten RJPM, Bouter LM. Manual therapy, physical therapy or continued care by a general practitioner for patients with neck pain. A randomized controlled trial. Ann Int Med 2002; 136: 713-722.
14. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, Bouter LM, de Vet HCW. The COSMIN Checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Medical Research Methodology 2010;10:22.

15. Ailliet L, Rubinstein SM, de Vet HCW. Characteristics of chiropractors and their patients in Belgium. J Manipulative Physiol Ther 2010; 33: 618-625.
16. Muthén LK, Muthén BO. (1998-2010). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.
17. Hu LT, Bentler PM. Cutoff criteria for fit indices in covariance structure analysis: conventional versus new alternatives. Structural equation modeling 1999;6(1):1-55
18. Muraki E. A generalized partial credit model: application of an EM algorithm. Applied Psychological Measurement 1992;16(2):159-176.
19. Wijnhoven HA, de Vet HC, Picavet HS. Explaining sex differences in chronic musculoskeletal pain in a general population. Pain 2006;124(1-2):158-166.
20. Glas CAW. Detection of differential item functioning using Lagrange multiplier tests. Statistica Sinica 1998;8(3):647-667.
21. van Groen MM, ten Klooster PM, Taal E, van de Laar MRT, Glas CAW. Application of the health assessment questionnaire disability index to various rheumatic diseases. Qual Life Res 2010;19:1255-1263.
22. Glas CAW. Modification indices for the 2-PL and the nominal response model. Psychometrika 1999;64(3):273-294.
23. Glas CAW. Preliminary Manual of the software program Multidimensional Item Response Theory (MIRT) 2010, Enschede, the Netherlands: University of Twente.
24. Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand. The Upper Extremity Collaborative Group (UECG). Am J Ind Med 1996 Jun;29(6):602-608.
25. Nordin M, Carragee EJ, Hogg-Johnson S, Schechter Weiner S, Hurwitz EL, Peloso PM, Guzman J, van der Velde G, Carroll LJ, Holm LW, Côté P, Cassidy DJ, Haldemna S. Assessment of neck pain and its associated disorders. Spine 2008; 33 (4S): S101-S122.
26. Soohoo NF, McDonald AP, Seiler JG, McGillivary GR. Evaluation of the construct validity of the DASH questionnaire by correlation to the SF-36. J Hand Surg Am 2002;27:537-541.
27. Bot SDM, Terwee CB, van der Windt DAWM, Bouter LM, Dekker J, de Vet HCW. Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. Ann Rheum Dis 2004;63:335-341.
28. Hains F., Waalen J., Mior S. Psychometric properties of the Neck Disability Index. J. Manipulative Physiol Ther 1998; 21: 75-80.
29. Verbrugge LM, Jette AM. The disablement process. Soc. Sci. Med. 1994;38(1):1-14.
30. Chok B, Gomez E. The reliability and application of the Neck Disability Index in physiotherapy. Physiother Singapore 2003;3:16-19.
31. van der Velde G, Beaton D, Hoog-Johnston S, Hurwitz E, Tennant A. Rasch analysis provides new insight into the measurement properties of the Neck Disability Index. Arthritis & Rheumatism 2009;61(4):544-551.

32. Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary physical function item bank supports the expected advantages of the Patient–Reported Outcomes Measurement Information System (PROMIS). Journal of Clinical Epidemiology 2008;61(1):17–33.
33. Amtmann DA, Cook KF, Jensen MP, Chen WH, Choi SW, Revicki D, Cella D, Rothrock N, Keefe F, Callahan L, Lai JS. Development of a PROMIS item bank to measure pain interference. Pain 2010;150(1):173–178.
34. US Food and Drug Administration Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. Rockville, MD: Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, 2009.

Neck Disability Index (NDI)

Vernon H and Hagino C, 1991 (with permission from Fairbank J)

1 *Pijn*

A. Ik heb nu geen pijn.
B. Ik heb nu weinig pijn.
C. Ik heb nu matige pijn.
D. Ik heb nu vrij hevige pijn.
E. Ik heb nu zeer hevige pijn.
F. Ik heb nu de slechts denkbare pijn.

2 *Persoonlijke verzorging (wassen, aan- en uitkleden))*

A. Ik kan goed voor mezelf zorgen zonder dat de pijn toeneemt.
B. Ik kan goed voor mezelf zorgen hoewel dat de pijn doet toenemen.
C. Voor mezelf zorgen is pijnlijk en gaat langzaam en voorzichtig.
D. Voor mezelf zorgen lukt goed maar vaak met enige hulp.
E. Elke dag voor mezelf zorgen lukt meestal alleen met hulp.
F. Ik kan mezelf niet aankleden; mezelf wassen gaat moeilijk en ik blijf in bed.

3 *Tillen*

A. Ik kan een zwaar gewicht tillen zonder dat de pijn toeneemt.
B. Ik kan een zwaar gewicht tillen, maar dat doet de pijn toenemen.
C. De pijn weerhoudt mij van het optillen van een zwaar gewicht van de grond, maar zou dat wel kunnen wanneer dat gewicht hoger (bijv. op een tafel) gelegen is.
D. De pijn weerhoudt mij ervan om zware dingen op te tillen, maar het lukt me wel om lichte tot middelzware gewichten te tillen als ze makkelijk geplaatst zijn.
E. Ik kan alleen zeer lichte gewichten tillen.
F. Ik kan helemaal niets tillen of dragen.

4 *Lezen*

A. Ik kan zo veel lezen als ik wil zonder pijn in mijn nek.
B. Ik kan zo veel lezen als ik wil met weinig pijn in mijn nek.
C. Ik kan zo veel lezen als ik wil met matige pijn in mijn nek.
D. Ik kan niet zo veel lezen als ik zou willen vanwege de matige pijn in mijn nek.
E. Ik kan bijna niet meer lezen vanwege de hevige pijn in mijn nek.
F. Ik kan helemaal niet meer lezen.

5 *Hoofdpijn*

A. Ik heb helemaal geen hoofdpijn.
B. Ik heb af en toe lichte hoofdpijn.
C. Ik heb af en toe matige hoofdpijn.
D. Ik heb vaak matige hoofdpijn.
E. Ik heb vaak hevige hoofdpijn.
F. Ik heb bijna altijd hoofdpijn.

6 *Concentratie*

A. Ik kan mij goed concentreren zonder moeite wanneer ik dat wil.
B. Ik kan mij goed concentreren met enige moeite wanneer ik dat wil.
C. Het kost mij duidelijk moeite om te concentreren wanneer ik dat wil.
D. Het kost mij veel moeite om te concentreren wanneer ik dat wil.
E. Het kost mij zeer veel moeite om te concentreren wanneer ik dat wil.
F. Ik kan mij helemaal niet concentreren.

7 *Werk*

A. Ik kan zo veel werk doen als ik wil.
B. Ik kan alleen mijn gewone werk doen, maar niet meer.
C. Ik kan het grootste deel van mijn gewone werk doen, maar niet meer.
D. Ik kan mijn gewone werk niet doen.
E. Ik kan bijna geen enkel werk meer doen.
F. Ik kan helemaal niet meer werken.

8 *Autorijden*

A. Ik kan autorijden zonder enige nekpijn.
B. Ik kan autorijden zo lang als ik wil met weinig pijn in mijn nek.
C. Ik kan autorijden zo lang als ik wil met matige pijn in mijn nek.
D. Ik kan niet autorijden zo lang als ik wil vanwege de matige pijn in mijn nek.
E. Ik kan bijna niet meer autorijden vanwege de hevige pijn in mijn nek.
F. Ik kan helemaal niet meer autorijden.

9 *Slapen*

A. Ik heb geen moeite met slapen
B. Mijn slaap is heel licht gestoord (minder dan 1 uur wakker).
C. Mijn slaap is licht gestoord (1 tot 2 uur wakker).
D. Mijn slaap is matig gestoord (2 tot 3 uur wakker).
E. Mijn slaap is fors gestoord (3 tot 5 uur wakker).
F. Mijn slaap is volledig gestoord (5 tot 7 uur wakker).

10 *Vrije tijd*

A. Ik kan aan alle activiteiten meedoen zonder enige pijn in mijn nek.
B. Ik kan aan alle activiteiten meedoen met enige pijn in mijn nek.
C. Vanwege de pijn in mijn nek kan ik aan de meeste, maar niet alle, gebruikelijke activiteiten meedoen.
D. Vanwege de pijn in mijn nek kan ik aan maar weinig gebruikelijke activiteiten meedoen.
E. Vanwege de pijn in mijn nek kan ik nagenoeg aan geen activiteiten meedoen.
F. Ik kan aan geen enkele activiteit meer meedoen.

SIGNATURE: ………………………………………………………… DATE: …………………………………………………………

DISABILITY INDEX SCORE: …………… %

INTERPRETATIE

Per vraag zijn er 6 antwoordcategorieën. De eerste antwoordcategorie (score 0) geeft geen beperkingen aan, de laatste categorie (score 5) betekent de meeste beperkingen. De totaalscore is de som van de tien delen vragen (maximaal 50) vermenigvuldigd met factor 2. De gevonden waarde representeert het "beperkingen-percentage" (0-100%).