J Rehabil Med 2015; 47: 836–843

ORIGINAL REPORT

RASCH ANALYSIS OF THE NECK BOURNEMOUTH QUESTIONNAIRE TO MEASURE DISABILITY RELATED TO CHRONIC NECK PAIN

Tommaso Geri, PT, MSc[1], Daniele Piscitelli, PT, MSc[2], Roberto Meroni, PT, PhD[2], Francesca Bonetti, PT[3], Giuseppe Giovannico, PT, MSc[3], Roberto Traversi, MD[4] and Marco Testa, PT, DO[1]

From the [1]Department of Neuroscience, Rehabilitation, Ophthalmology, Genetics, Maternal and Child Health, University of Genova, Campus of Savona, [2]Department of Translational medicine and Surgery, University of Milano-Bicocca, [3]Department of Clinical Sciences and Translational Medicine, University of Rome - Tor Vergata and [4]Italian Rehabilitation and Reintegration of Invalids Association (AIRRI), Rome Clodio, Italy

Objective: **To determine the psychometric properties of the Neck Bournemouth Questionnaire in patients with chronic neck pain, using Rasch analysis.**

Methods: **A sample of 161 subjects with chronic neck pain was assessed with the Neck Bournemouth Questionnaire. Before performing Rasch analysis, we examined the structure of the scale with factor analysis. The goodness-of-fit of the data to the model, thresholds ordering, unidimensionality, local independence of the items, differential item functioning, person separation index, and mean person's location were assessed.**

Results: **Both exploratory and confirmatory factor analyses supported the presence of 2 factors. Only Factor 1 needed a modification (item 7 removal) in order to achieve the fit to the Rasch model (χ^2=10.65, df 8, p=0.22). The person separation index was 0.80 and the mean location of persons 0.48 (standard deviation (SD) 1.02). Factor 2 (items 4 and 5) fitted the model without modifications (χ^2=3.86, df 4, p=0.42). Its person separation index and mean person's location were, respectively, 0.77 and –0.71 (SD 1.57).**

Conclusion: **The Neck Bournemouth Questionnaire with the purposed modification may provide useful clinical profiles and change scores of subjects with chronic neck pain for research purposes.**

Key words: neck pain; outcome assessment; validation studies; Rasch analysis; Neck Bournemouth Questionnaire.

J Rehabil Med 2015; 47: 836–843

Correspondence address: Tommaso Geri, Department of Neuroscience, Rehabilitation, Ophthalmology, Genetics, Maternal and Child Health, University of Genova, Campus of Savona, Via A. Magliotto 2, IT-17100 Savona, Italy. E-mail: tommaso.geri@gmail.com

Accepted May 20, 2015; Epub ahead of print Jul 16, 2015

INTRODUCTION

Neck pain (NP) contributes significantly to the burden of global disability (1), because its prevalence is high and many people with NP experience activity limitations (2). Among the patient-reported outcomes measures regarding NP, 2 recent systematic reviews critically appraised the original (3) and translated (4) versions of neck-specific questionnaires using well-accepted methodological standards (5). Both recommended the use of the Neck Disability Index (NDI) (6), because of its well-validated properties supported by positive findings. Thereby, several arguments suggest that the NDI might be reconsidered as the instrument of first-choice. The content validity of the NDI has been questioned because of discrepancies in the nature of the measured construct (7). Furthermore, the NDI lacks unidimensionality (7–9). In a multidimensional ordinal scale, the sum score of the entire scale is inappropriate to calculate change scores and to analyse using parametric statistics, as it cannot be considered an interval-level scale (10). Moreover, the presence of a large floor-effect and a marginal ceiling-effect (9) may mean that the NDI is unable to capture the clinical status of people with extreme scores. Finally, the reliability of the NDI seems inadequate (3). Therefore, different neck-specific questionnaires, which may overcome the limits of the NDI, are needed (3, 4, 7–9).

A systematic review (11) pointed out that, when it comes to the ICF category coverage of a disease-specific questionnaire for NP, the most inclusive were the NDI (6), the Neck Pain and Disability Scale (12) and the Neck Bournemouth Questionnaire (NBQ) (13). The NBQ was adapted by Bolton & Humphreys (13) from the Bournemouth Questionnaire, initially developed in English for patients with low-back pain. The NBQ has been shown to have acceptable psychometric properties for use in clinical and research settings (13) and it has been translated and validated in German (14), Dutch (15), French (16) and Italian (17). Recent evidence regarding its internal consistency (14, 17), content validity (17, 18), structural validity (17) and interpretability (17) are promising, even though they still need to be critically evaluated. Instead, evidence reporting that the NBQ has high reliability (15), substantial responsiveness and construct validity (13) are considered limited because these positive results have been found in studies of fair methodological quality (3, 4). As a proper validation of reliability and responsiveness requires interval scores, the assumption of unidimensionality, which is a fundamental requirement of construct validity (19), needs to be tested.

Among modern psychometric approaches, Rasch analysis is a method developed to test whether scores on items from a

 doi: 10.2340/16501977-2001
 ISSN 1650-1977

questionnaire fit a measurement model considered fundamental to assume scores as an interval variable and as invariant across relevant demographic and clinical characteristics (19, 20). Therefore, the aim of this study was to use Rasch analysis to test whether the NBQ fits the assumptions of unidimensional construct, local independence and absence of differential item functioning (DIF) (a detailed explanation of these terms is provided in the Methods section), which are mandatory in order to consider the NBQ an interval-level scale.

METHODS

The analysis was performed with the Italian version of the NBQ (NBQ-I) (17). SPSS release 21.0 was used to perform the descriptive analyses and exploratory factor analysis. Confirmatory factor analysis was performed using IBM-SPSS Amos-22. RUMM2030 (21) was used for Rasch analysis.

Patients and setting

A convenience sample of 161 patients was recruited from patients attending the outpatient physiotherapy services of the Santa Corona Hospital (Pietra Ligure, Italy) and affiliated centres from September 2012 to February 2014. Inclusion criteria were: a diagnosis of chronic non-specific NP (> 3 months) provided by a medical doctor; age > 18 years; and the ability to read and speak Italian fluently. Exclusion criteria were: specific NP; psychiatric and mental deficits; central or peripheral neurological signs; systemic illness; clinical instability (cardiac, respiratory, vascular) and vertebral surgery. A sample size of 150 subjects was estimated to achieve an item calibration stability of ±½ logit with a confidence interval of 99% (22). All subjects gave their informed consent prior to participating in the study, which received the approval of the ethics committee of the Azienda Sanitaria Locale 2 Savonese (no. 650–04/07/2013).

Scale description

The NBQ consists of 7 items dealing with: level of pain, physical function, social activity, anxiety, depression, work-related fear avoidance beliefs, and coping strategies for pain control. Each item has 11 numerical response categories, ranging from 0 to 10, where 0 corresponds to the absence of the problem and 10 to the highest level of the problem mentioned in the corresponding item. The total raw score ranges between 0 and 70 and is calculated as the sum of the raw scores of each item.

Structural validity

As the Rasch model assumes that all items of a scale are measuring a single underlying construct, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were performed to assess the unidimensionality of the NBQ-I prior to continue with Rasch modelling.

A significant ($p < 0.05$) Bartlett's test of sphericity, which examines whether the correlations in the data-set are appropriate, and a Kaiser-Meyer-Olkin Measure (KMO) > 0.80, which tests the sampling adequacy to ensure that the scale items are relevant, were the criteria considered before performing EFA. EFA was performed by a maximum likelihood (ML) extraction method. Factors were extracted when their eigenvalues were > 1 and using the Cattell's Scree Test. The amount of variance explained by the extracted factors was reported. Item factor loadings were obtained using Promax Rotation with Kaiser normalization. Each item was considered as belonging to the factor when its load was > 0.40 (23), and communalities were reported. A CFA was subsequently performed, on the basis of EFA, by means of the ML extraction method. The model fit was evaluated using the following fit indices: the χ^2 test, which compares the fitted model with the saturated model that fits the covariance perfectly and indicates a good fit when it is not significant; the ratio between the χ^2 test and degrees of freedom (χ^2 test/df), considered as a good fit when the ratio is < 3; the root mean square error of approximation (RMSEA) with 90% confidence interval (CI), that should be close to zero and ideally less than 0.05; the comparative fit index (CFI) and the Tucker–Lewis Index (TLI), that were considered indicative of good model fit for values higher than 0.90 and 0.95, respectively (24). If unidimensionality was not confirmed, then the number of underlying factors and their relation with each item were used to separate the scale into subscales, each subjected to Rasch analysis.

Rasch analysis

Rasch analysis was performed to study the fit of the (sub)scales to the Rasch model. A detailed explanation of the model and processes can be found in Rasch (25) and Tesio (26).

The fit statistics used to test whether data satisfied the model's expectations were as follows:

- Initially, a likelihood ratio test was conducted to determine whether to use the Rating Scale Model (RSM) or the Partial Credit Model (PCM). A statistically significant ($p < 0.05$) likelihood ratio test suggests the use of PCM because the distance between options is not uniform (20).
- Goodness-of-fit statistics were applied using a χ^2 interaction statistic for both overall data and individual items and persons. The overall data fit indicates no substantial deviation from the model when χ^2 value is non-significant ($p > 0.05$). Individual item and person fits were tested considering their fit residuals, the differences between the actual values and the values estimated from the model. Adequate individual fits were achieved if the χ^2 value was not significant ($p > 0.05$) and the standardized fit residuals were between –2.5 and +2.5. Fit residuals above 2.5 indicate the item measures a construct different from other items, whereas fit residuals below –2.5 indicate redundancy. Furthermore, fit residuals of items and persons should have the mean value close to 0 and standard deviation (SD) approximately 1.0 (20). The significance level was corrected according to the Bonferroni method. Subjects with extreme scores, i.e. the maximum or the minimum scores of the scale, were excluded from the modelling as they do not contain any information for estimating person parameters (27).
- Item threshold ordering was considered as each item of the NBQ-I has 11 levels of agreement with the item statement, and thus 10 thresholds, which increase consistently with the underlying trait. If the thresholds were disordered, adjacent response categories were collapsed (20).
- Local independence assumes that no significant association among item responses should be found once the dominant factor influencing a person's response to those items has been conditioned out. Local independence of the items was confirmed when the residual correlation matrix of the items had values < 0.30 (28). The unidimensionality assumption was then evaluated with the test proposed by Smith (29), which uses an independent *t*-test procedure to assess differences between person estimates, derived from 2 subsets of items (of the same scale) identified by positive and negative loadings on the first principal component of the residuals. The scale was considered unidimensional when less than 5% of the estimates are outside the range of ±1.96, or the lower bound of the binomial confidence interval overlaps 5%.
- The absence of DIF ensures the invariability of the measure (30). It was examined with an analysis of variance, with a Bonferroni correction for each level of each variable, in which the responses to a single item are compared across each levels of subject characteristics, referred to as person factor, and across different levels of latent trait, referred to as class intervals. We tested DIF for the person factor age (3 levels: under 45 years, 46–65 years, over 66 years), sex (2 levels: male, female), educational level (3 levels: elementary-middle school, high school, university), marital status (2 levels: yes, no), working status (3 level: employed, unemployed, retired) and current smoking (2 levels: yes, no).

Table I. *Demographic characteristics*

Variables	*n* (%)
Sex	
Female	116 (72.0)
Male	45 (28.0)
Marital status	
Yes	108 (67.1)
No	53 (32.9)
Educational level	
Elementary – Middle	34 (21.1)
High school	76 (47.2)
University	51 (31.7)
Work	
Student	5 (3.1)
Employed	83 (51.6)
Self-employed	25 (15.5)
Retired	29 (18.0)
Unemployed	5 (3.1)
Housewife	14 (8.7)
Smoking	
Yes	36 (22.4)
No	125 (77.6)
Radiation of symptoms	
Yes	72 (44.7)
No	89 (55.3)

- The presence of uniform DIF was identified when person factor levels significantly influence the item responses, whereas a non-uniform DIF was detected when the interaction between person factor and class interval significantly alters the item responses. Further, the item characteristic curves were examined. A split procedure, which consists in dividing (splitting) the item presenting a DIF for the person factors (31), was considered if an item exhibited a uniform DIF. The relevance on person estimate of uniform DIF was examined with an anchored analysis, which compares person estimates derived from an "item pure set" (deleting items displaying uniform DIF) and from the "original full set" anchored to the pure set's items parameter estimates (item difficulties + Rasch-Andrich thresholds) (30).
- Item deletion was applied on items having non-uniform DIF because their variance, which varies across ability levels, cannot be adjusted (31).
- Scale targeting was assessed by the inspection of the person item-threshold distribution that compares the distribution of patients' location with the location of the thresholds along the same metric logit scale. A mean (SD) location of person approximating the values 0 (1) indicated a good targeting. The mean location of item difficulty was set by default at 0 logit (20). Floor- and ceiling-effects were considered present when more than 15% of the subjects reached, respectively, the lowest and the highest scores.
- The person separation index (PSI), which is a measure of internal consistency of the scale, gives an indication of the power of the scale to discriminate among persons with different levels of the trait. A value of 0.7 is considered a minimal value for group or research use and of 0.85 for individual or clinical use (20).

Table II. *Result of exploratory factor analysis*

Item	Factor 1	Factor 2	Communalities
1	**0.792**	–0.012	0.616
2	**1.009**	–0.085	0.924
3	**0.863**	0.010	0.754
4	0.114	**0.867**	0.880
5	–0.053	**0.881**	0.723
6	**0.604**	0.168	0.512
7	**0.424**	0.168	0.292

Loading values on each factor are in bold.

If all the above-mentioned fit statistics were acceptable, then the data fit the Rasch model and, accordingly, ordinal raw scores were converted into interval scores.

RESULTS

Subjects

There were 161 participants, of whom 116 (72%) were women. The mean age was 49.9 years (SD 13.4, range 24–79 years), BMI mean was 24 kg/m^2 (SD 3.85, range 15.6–38.9 kg/m^2). Demographic characteristics of the sample are shown in Table I. The NBQ-I had a median of 27 (range 3–69).

Structural validity

Bartlett's test was significant ($p < 0.05$) and the KMO was 0.83. Therefore, we proceeded with EFA. Two factors with eigenvalues >1 explained 67.2% of total variance. The first factor explained 55.9% and the second 11.3%. The number of factors with eigenvalue > 1 and the Cattel's Scree Plot confirmed the 2-factor structure. Rotation of the solution revealed the items loadings and communalities across the 2 factors (Table II). The communalities highlighted that neither F1 nor F2 explained a substantial proportion of variance of item 7. The correlation between the 2 factors was 0.59. The CFA confirmed the 2 factors structure as all fit indices suggested an adequate model fit. The χ^2 was not significant ($\chi^2 = 11.727$, df 11, $p = 0.38$), the RMSEA was 0.02 (90% CI 0.00–0.087; test for RMSEA ≤ 0.05 (p-value) = 0.69], the ratio χ^2 test/df was 1.07, and the CFI and TLI were, respectively, 0.999 and 0.998.

Rasch analysis on NBQ-I/F1 (items 1, 2, 3, 6, 7)

Significant likelihood ratio tests ($p < 0.05$) supported the use of PCM. The overall fit statistics indicated significant misfit to the model ($\chi^2 = 37.6119$, df 10, $p < 0.01$). The item- and person-fit residuals had, respectively, means of 0.1 (SD 1.66) and –0.55 (SD 1.27) (Table III). Three subjects with extreme

Table III. *Summary of Rasch analyses of the Neck Bournemouth Questionnaire*

	Overall fit statistics		Item-fit residuals,	Person-fit residual,		Test of unidimensionality
Analysis name	χ^2 (df)	p-value	Mean (SD)	Mean (SD)	PSI	percent (95% CI)
NBQ-I Factor 1 (items 1, 2, 3, 6, 7)	37.61 (10)	<0.001	0.10 (1.66)	–0.55 (1.27)	0.82	6.83 (3.86–11.82)
NBQ-I Factor 1 without item 7	10.65 (8)	0.22	–0.15 (1.57)	–0.20 (0.94)	0.80	2.48 (0.97–6.20)
NBQ-I Factor 2 (items 4, 5)	3.87 (4)	0.42	0.11 (0.24)	–0.40 (0.68)	0.77	2.48 (0.97–6.20)

NBQ-I: Neck Bournemouth Questionnaire – Italian version; CI: confidence interval; df: degrees of freedom; PSI: Person Separation Index; SD: standard deviation.

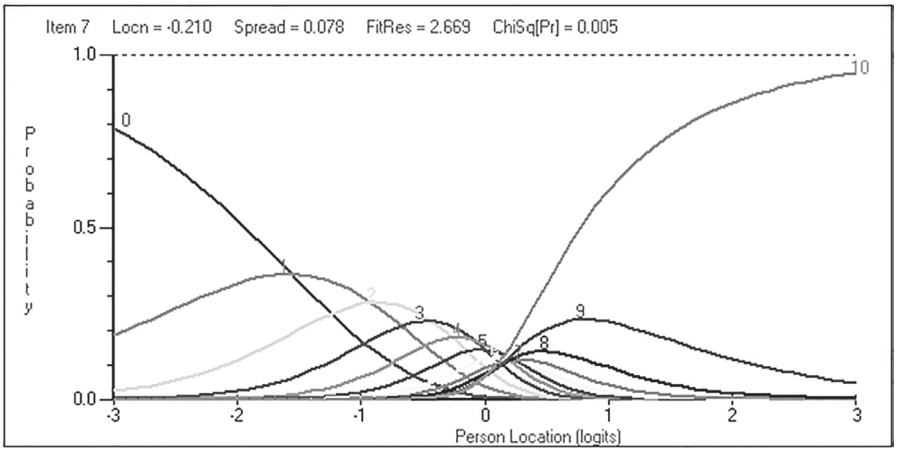

Fig. 1. Category probability curves of item 7, illustrates transitions between response categories. The y-axis represents the probability of the category answer. The x-axis represents the logits. Locn: location; FitRes: fit residual; ChiSq: chi-square; Pr: probability value; F: F statistic.

scores were discarded from this analysis (2 had the minimum and 1 the maximum possible scores). Ten subjects with fit residuals below –2.5 did not fit the model.

Item 7, which asks whether "Over the past week, how much have you been able to control (reduce/help) your neck pain on your own?", exhibited misfit (fit residual = 2.669, $\chi^2 = 10.775$, df 2, $p = 0.01$).

All the items had ordered thresholds except item 7, whose categories from 6 to 9 did not emerge (Fig. 1). The comparison of 2 subsets of the items indicated multidimensionality as 6.83% (95% CI = 3.86–11.82%) of the total tests were significant. No evidence of DIF or response dependency was found. As Item 7 showed disordered thresholds, it was re-scored with the aim of adjusting its misfit by collapsing categories according to a pattern of 0, 1, 2, 3, 4, 5, 5, 6, 6, 7, and 7. This procedure did not improve the item-fit residual (fit residual = 2.677, $\chi^2 = 10.235$, df 2, $p = 0.01$) and the overall fit statistic still indicated misfit ($\chi^2 = 26.8403$, df 10, $p = 0.01$).

Therefore, item 7 was removed from the analysis because: (*i*) it demonstrated a low load on both factors with low communalities values (< 0.3); (*ii*) it misfitted the model during the first analysis (with a disordered categories pattern) and after the rescoring procedure; (*iii*) the *t*-test procedure indicated multidimensionality; and (*iv*) its wording, which relates to coping strategies, makes it conceptually different from factor 1, which deals with pain and functioning.

Following removal of item 7, a satisfactory fit to the model was achieved ($\chi^2 = 10.65$, df 8, $p = 0.22$). The item-fit residuals had mean –0.15 (SD 1.57) and the person-fit residuals had mean –0.20 (SD 0.94) (Table III). All individual item-fit showed normal fit residuals value (Table IV) while 5 subjects had individual person-fit residuals below –2.5. Four subjects with extreme scores were discarded from this analysis (3 had the minimum and 1 the maximum possible scores). All the items had ordered thresholds. The local independence of the items was confirmed as no pairs of items exceeded the threshold value after exploring the residual correlation matrix. Furthermore, the independent *t*-test procedure resulted in 4 significant tests out of 161 (2.48%, 95% CI 0.97–6.2%) and therefore confirmed the unidimensionality of NBQ-I/F1 (without item 7). A uniform DIF for educational level ($F = 6.84599$, df 2, $p = 0.01$) was found for item 2 as the Elementary-Middle School population scored higher than expected. A slight improvement of the overall fit of the scale was noted ($\chi^2 = 12.2578$, df 10, $p = 0.27$) after splitting item 2 into

Table IV. *Individual item-fit of the 2 factors of the Neck Bournemouth Questionnaire (NBQ)*

Factor	Item	Question	Location[a]	SE	Fit Res	χ^2 (df)	p^b
Factor 1	1	Over the past week, on average how would you rate your neck pain?	–0.29	0.05	0.35	0.08 (2)	0.96
	2	Over the past week, how much has your neck pain interfered with your daily activities (housework, washing, dressing, lifting, reading, driving)?	0.17	0.05	–0.98	5.04 (2)	0.08
	3	Over the past week, how much has your neck pain interfered with your ability to take part in recreational, social and family activities?	0.32	0.05	–1.13	3.02 (2)	0.22
	6	Over the past week, how have you felt your work (both inside and outside the home) has affected or would affect your neck pain?	–0.20	0.05	2.04	2.46 (2)	0.29
Factor 2	4	Over the past week, how anxious (tense, uptight, irritable, difficulty in concentrating/relaxing) have you been feeling?	–0.22	0.06	–0.06	2.25 (2)	0.32
	5	Over the past week, how depressed (down-in-the-dumps, sad, in low spirits, pessimistic, unhappy) have you been feeling?	0.22	0.06	0.28	1.62 (2)	0.44

[a]Location, item difficulty expressed in logits.
[b]Bonferroni-corrected χ^2 *p*-value was applied.
df: degrees of freedom; *p*: probability value; SE: standard error.

Elementary–Middle School and High School–University. The uniform DIF was not relevant on person estimate. Only a subject shown a person estimate greater than 0.5 logit, between subjects estimate derived from the pure "set" and the anchored full set. Non-uniform DIFs were not detected in any item.

The targeting of NBQ-I/F1 (without item 7) showed 3 (1.86%) and 1 subjects (0.62%) located, respectively, at the floor and the ceiling of the subscale. The mean location of persons was –0.48 logit (SD 1.02) (Fig. 2A). The PSI was 0.80 (Table III).

The conversion table of NBQ-I/Subscale-1, which allows the transformation of raw scores into interval scores, is presented in Table V. The relationship between ordinal and interval scores of the scale is presented in Fig. 3.

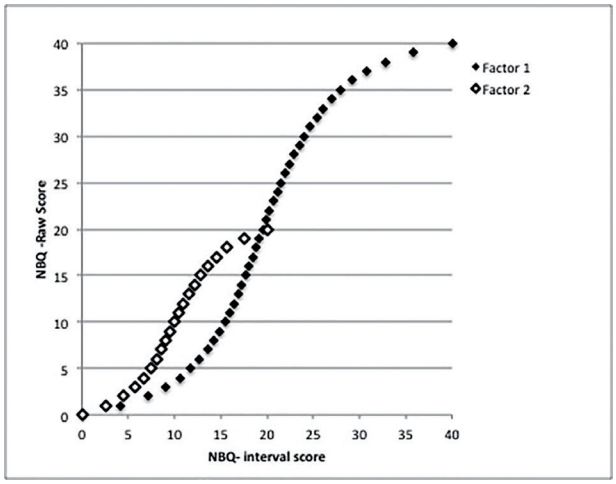

Fig. 3. Conversion curves from ordinal to interval scores of Neck Bournemouth Questionnaire – Italian version (NBQ-I)/Subscale 1 (*black squares*) and NBQ-I/Subscale 2 (*white squares*). Note: item 7 was deleted in Subscale 1.

Rasch analysis on NBQ-I/F2 (items 4, 5)

Likelihood ratio test was significant ($p < 0.05$), thus a PCM was applied. The subscale showed good overall fit to the model ($\chi^2 = 3.8658$, df 4, $p = 0.42$). The mean item- and person-fit residuals were, respectively, 0.10 (SD 0.24) and –0.40 (SD 0.68) (Table III), with no misfit observed for items (Table IV) and persons. All items presented with ordered thresholds. The local independence of the items was confirmed as the analysis of response dependency showed correlations < 0.3 and the assessment of unidimensionality resulted in 4 (2.48% 95% CI =0.97–6.2%) significant *t*-tests. No DIFs were found across the tested person factors and class intervals. The targeting of NBQ-I/F2 showed the presence of 17 (10.6%) and 3 (1.9%) persons at, respectively, the floor and the ceiling of the sub-scale (Fig. 2B). The mean person location was –0.71 logit (SD 1.57). The PSI was 0.77. The conversion table of

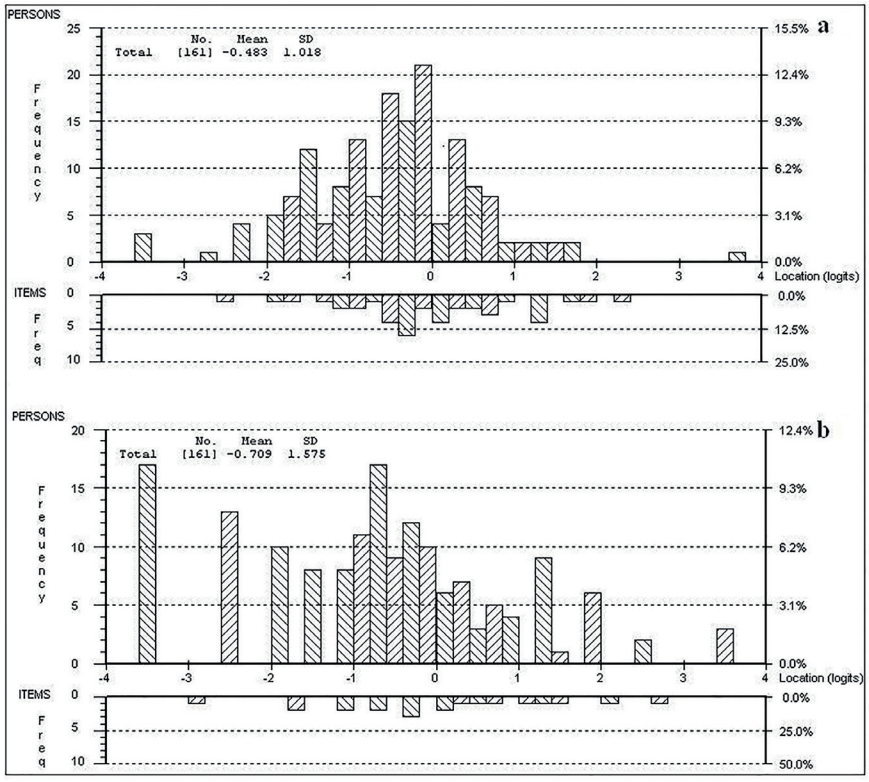

Fig. 2. Distribution of the items and patients ($n = 161$) along the Rasch-calibrated metric scale. (a and b) *Upper panel* shows the location of the subjects. *Lower panel* shows the threshold of the items. (a) Referred to Neck Bournemouth Questionnaire – Italian version (NBQ-I) factor 1. (b) Referred to NBQ-I factor 2. Grouping set to interval length of 0.20, making 40 groups.

Table V. *Conversion table between Neck Bournemouth Questionnaire – Italian version (NBQ-I)/Subscale 1 (S1) ordinal (summative) scores and interval scores*

NBQ-I/S1 raw score[a]	Logit location	NBQ-I/S1 interval score (0–40)[b]
0	–3.54	0.00
1	–2.79	4.19
2	–2.26	7.15
3	–1.91	9.12
4	–1.64	10.60
5	–1.43	11.77
6	–1.26	12.74
7	–1.11	13.57
8	–0.98	14.28
9	–0.87	14.91
10	–0.77	15.47
11	–0.68	15.97
12	–0.59	16.44
13	–0.52	16.87
14	–0.44	17.28
15	–0.37	17.68
16	–0.31	18.05
17	–0.24	18.42
18	–0.17	18.79
19	–0.11	19.16
20	–0.04	19.53
21	0.03	19.91
22	0.10	20.30
23	0.17	20.69
24	0.24	21.10
25	0.32	21.53
26	0.40	21.97
27	0.48	22.44
28	0.57	22.94
29	0.67	23.47
30	0.77	24.04
31	0.88	24.66
32	1.00	25.35
33	1.14	26.12
34	1.30	27.00
35	1.48	28.02
36	1.70	29.24
37	1.97	30.75
38	2.33	32.76
39	2.87	35.76
40	3.63	40.00

[a]Ordinal scores are obtained by summing raw responses of the 4 items (1, 2, 3, 6) of NBQ-I/S1.
[b]NBQ-I/S1 ordinal scores are transformed into a 0–40 interval scale. This conversion can be used only with the modified NBQ-I factor 1, without item 7.

Table VI. *Conversion table between Neck Bournemouth Questionnaire – Italian version (NBQ-I)/Subscale 2 (S2) ordinal (summative) scores and interval scores*

NBQ-I/S2 raw score[a]	Logit location	NBQ-I/S2 interval score (0–20)[b]
0	–3.45	0.00
1	–2.55	2.63
2	–1.89	4.55
3	–1.46	5.81
4	–1.15	6.73
5	–0.90	7.45
6	–0.70	8.04
7	–0.52	8.57
8	–0.35	9.05
9	–0.19	9.52
10	–0.03	9.99
11	0.14	10.48
12	0.31	10.99
13	0.50	11.55
14	0.71	12.15
15	0.94	12.83
16	1.20	13.60
17	1.52	14.53
18	1.93	15.73
19	2.54	17.51
20	3.39	20.00

[a]Ordinal scores are obtained by summing raw responses of the 2 items (4, 5) of NBQ-I/S2.
[b]NBQ-I/S2 ordinal scores are transformed into a 0–20 interval scale.

scores of subscale 2 is presented in Table VI and its structure is reported in Fig. 3.

DISCUSSION

This study provides additional support for the validity of the NBQ through the use of Rasch analysis, a modern psychometric approach that has previously disputed the validity of the NDI (7–9) in measuring health-related quality of life of people with NP. The structural validity, tested with both EFA and CFA, revealed that the NBQ-I has 2 factors, which were separately subjected to Rasch analysis. Subscale 1, which deals with pain and functioning (17), became an interval scale after removal of item 7, while Subscale 2, which deals with anxiety and depression (17), achieved interval-level scaling without adjustment. The uniform DIF due to educational level (within a minimal difference between high- and low-educational levels) found for item 2 may bias the scores of Subscale 1. However, as epidemiological data suggest a higher prevalence of NP disability in women, middle-aged subjects and workers (2), we did not found any DIF accounting for sex, age and employment status.

The targeting of the NBQ-I/Subscale 1 indicated that, on average, the difficulty of the items targeted the ability of the sample better than the NDI, as reported by van der Velde et al. (8), who found a mean location of person of –1.69 (SD 1.04), Walton & MacDermid (32), who had an analogous person-item threshold distribution assessed visually, and Johansen et al. (33), who reported a mean location of person of –1.17 (SD 1.39). Moreover, we observed the absence of floor- and ceiling-effects for Subscale 1, while a large floor-effect has been reported for the NDI (9). This aspect further confirms that the NDI is unable to measure the lower levels of neck-pain-related disability. Furthermore, the PSI of Subscale 1 allows its use for research purpose. Previous studies on the NDI (8, 32) reported PSI values similar to the values obtained in this study. Despite the fact that these data are from different samples, our sample was comparable in terms of male/female ratio, mean age and pain intensity. Therefore, it might be argued that the NBQ-I/Subscale 1, intended to measure a construct of neck-related disability similar to that of the NDI, better assesses the health status of the patients with chronic NP in research settings.

The aetiopathogenesis of NP is due to a complex interaction of physical, cognitive and environmental factors, whose combination plays an important role in determining the patient health status (34). High levels of baseline pain and self-reported neck disability have been recognized as predictors of prolonged recovery (35). Furthermore, it is also well established that psychological distress, characterized by depressed mood and increased anxiety, acts as predictor of prolonged recovery and may recognize those patients with high risk of transition to chronic pain (36). As NBQ-I/Subscale 1 and NBQ-I/Subscale 2 measure, respectively, baseline pain/self-reported disability and psychological distress, the NBQ-I is suitable to provide profiles of NP patients that may assist their classification for clinical decision-making purposes. However, the establishment of clinical pathways according to the profiles derived from the subscales of the NBQ-I still needs to be elucidated prospectively.

The examination of the NBQ-I with Rasch analysis provides a disease-specific questionnaire whose scores can be assumed as interval variables and, consequently, legitimately subjected to parametric statistics and interpreted as meaningful change scores in the assessment of treatment effectiveness. Indeed, even though the sum of raw scores is distributed normally, their logarithmic conversion into interval data gives a distribution curve (Fig. 3) similar to that observed for the NDI (8), where the majority of raw scores is concentrated at the mid-range of the scale, in which the slope of the distribution curve is the steepest, while the decreased slope in the 2 tails spreads the extreme scores. This would suggest that change is not consistent across the entire breadth of the scale, which ought to have considerable impact for key features of a tool, such as clinically important differences. The same amount of change in raw score has a different weight on the transformed interval-level score based on the overall score achieved, The change score located at the mid-range of the ordinal scale may be considered clinically remarkable, yet it may be small for clinicians and patients when intended at the interval-level, i.e. a raw score change of 5 points with an overall score of 20 (scale total score changes from 25 to 20) represents a change of 2 points after an interval-level transformation. Conversely, a change score located at the extreme of the ordinal scale may be considered clinically negligible despite the substantial distance between interval-level scores, i.e. the same raw score difference (5 points) with an overall score of 35 (scale total score changes from 40 to 35) represents a change of 11.98 points after an interval-level transformation (see Fig. 3 and Tables V for reference). As a consequence, the interpretation of change scores of raw scores may lead to biased conclusions about treatment effectiveness and responsiveness. Therefore, our recommendation is to use the conversion tables provided in the present study for the subscales of the NBQ-I, in order to achieve unbiased conclusions about treatment effectiveness when measuring disability in subjects with NP.

The present findings add to the body of knowledge on disease-specific questionnaires measuring disability related to chronic NP. Even though previous systematic reviews (3, 4) on the psychometric properties of patient-reported outcome measures concerning NP disability recommended the use of the NDI, they warranted for the evaluation of the unknown measurement properties of other questionnaires. As this study improved the construct validity of the NBQ, which seems superior to that of the NDI, we believe that further systematic reviews addressing patient-reported outcome instruments may consider the NBQ as a valid instrument for measuring health-related quality of life in people with chronic NP.

We suggest the following implementations for further psychometric validations of the NBQ. The interval scores obtained with the conversion tables, which apply for similar patients and settings, should be used to calculate change scores. For Subscale 1, the transformation into interval score was obtained after removal of item 7. As Linacre (37) suggested 10 subjects for response category to estimate the sample size of Rasch analysis studies for questionnaires with polytomous items, our sample may be considered modest to lead to a firm deletion of item 7. However, Chen et al. (38) reported a number greater than 100 subjects may lead to valid Rasch analysis. Despite the promising psychometric properties of the NBQ in the Italian population, we are aware that its usefulness would be extended in clinical and research settings worldwide only after validation of an English version, accounting whether removal of item 7 alters the structural and content validity of the NBQ, is performed with similar techniques in larger samples. Furthermore, there is the need to test the concurrent validity of the present modified version. Finally, even though the NBQ is claimed to have good content validity because its factors cover the ICF constructs of impairment, disability, participation and personal factors, it does not include the investigation of environmental factors (11, 18).

A limitation of this study is the presence of only 2 items in NBQ-I/Subscale 2. A factor comprised of merely 2 items is largely uninterpretable because a factor vector can be fit between any 2 items. The minimal number of 3 items should contribute to a factor (39). As a factor can be represented geometrically as a vector fitted to points (items) in n-dimensional space and its performance as the distance between the points and the vector, a straight line can always fit perfectly 2 points. Furthermore, the small size of our sample may have produced a type II error in detecting non-uniform DIFs, as this analysis was underpowered (40).

In conclusion, Rasch analysis of the NBQ pointed out the unidimensionality of its subscales, whose psychometric properties were acceptable in terms of structural validity and internal consistency. Although Subscale 2 should be used with caution, Subscale 1 may be used in research settings as a tool for measuring the effectiveness of treatments aiming at reducing pain and improving function in patients with NP, using the conversion table of raw scores into interval scores.

REFERENCES

1. Hoy D, March L, Woolf A, Blyth F, Brooks P, Smith E, et al. The global burden of neck pain: estimates from the global burden of disease 2010 study. Ann Rheum Dis 2014; 73: 1309–1315.
2. Hogg-Johnson S, van der Velde G, Carroll LJ, Holm LW, Cassidy JD, Guzman J, et al. The burden and determinants of neck pain

in the general population: results of the Bone and Joint Decade 2000–2010 Task Force on Neck Pain and Its Associated Disorders. Spine (Phila Pa 1976) 2008; 33: S39–S51.
3. 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 2012; 21: 659–670.
4. Schellingerhout JM, Heymans MW, Verhagen AP, de Vet HC, Koes BW, Terwee CB. Measurement properties of translated versions of neck-specific questionnaires: a systematic review. BMC Med Res Methodol 2011; 11: 87.
5. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010; 63: 737–745.
6. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther 1991; 14: 409–415.
7. Ailliet L, Knol DL, Rubinstein SM, de Vet HC, van Tulder MW, Terwee CB. Definition of the construct to be measured is a prerequisite for the assessment of validity. The Neck Disability Index as an example. J Clin Epidemiol 2013; 66: 775–782.
8. van der Velde G, Beaton D, Hogg-Johnston S, Hurwitz E, Tennant A. Rasch analysis provides new insights into the measurement properties of the neck disability index. Arthritis Rheum 2009; 61: 544–551.
9. Hung M, Cheng C, Hon SD, Franklin JD, Lawrence BD, Neese A, et al. Challenging the norm: further psychometric investigation of the neck disability index. Spine J 2014 Mar 22. [Epub ahead of print].
10. Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: time to end malpractice? J Rehabil Med 2012; 44: 97–98.
11. Ferreira ML, Borges BM, Rezende IL, Carvalho LP, Soares LP, Dabes RA, et al. Are neck pain scales and questionnaires compatible with the international classification of functioning, disability and health? A systematic review. Disabil Rehabil 2010; 32: 1539–1546.
12. Wheeler AH, Goolkasian P, Baird AC, Darden BV 2nd. Development of the Neck Pain and Disability Scale. Item analysis, face, and criterion-related validity. Spine (Phila Pa 1976) 1999; 24: 1290–1294.
13. Bolton JE, Humphreys BK. The Bournemouth Questionnaire: a short-form comprehensive outcome measure. II. Psychometric properties in neck pain patients. J Manipulative Physiol Ther 2002; 25: 141–148.
14. Soklic M, Peterson C, Humphreys BK. Translation and validation of the German version of the Bournemouth Questionnaire for Neck Pain. Chiropract Manual Ther 2012; 20: 2.
15. Schmitt MA, de Wijer A, van Genderen FR, van der Graaf Y, Helders PJ, van Meeteren NL. The Neck Bournemouth Questionnaire cross-cultural adaptation into Dutch and evaluation of its psychometric properties in a population with subacute and chronic whiplash associated disorders. Spine (Phila Pa 1976) 2009; 34: 2551–2561.
16. Martel J, Dugas C, Lafond D, Descarreaux M. Validation of the French version of the Bournemouth Questionnaire. J Can Chiropr Assoc 2009; 53: 102–120.
17. Geri T, Signori A, Gianola S, Rossettini G, Grenat G, Checchia G, et al. Cross-cultural adaptation and validation of the Neck Bournemouth Questionnaire in the Italian population. Qual Life Res 2015; 24: 735–745.
18. Schmitt MA, Schroder CD, Stenneberg MS, van Meeteren NL, Helders PJ, Pollard B, et al. Content validity of the Dutch version of the Neck Bournemouth Questionnaire. Man Ther 2013; 18: 386–389.
19. Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. Value Health 2004; 7 Suppl 1: S22–S26.
20. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum 2007; 57: 1358–1362.
21. Andrich D, Sheridan BE, Luo G. Manual for the Rasch Unidimensional Measurement Model (RUMM2030). Perth, Western Australia: RUMM Laboratory; 2010.
22. Linacre JM. Sample size and item calibration stability. Rasch Measurement Transaction 1994; 7: 328–331.
23. Stevens JP. Applied multivariate statistics for social sciences. 4th edn. Hillsdale: Lawrence Erlbaum; 2002.
24. Brown TA. Confirmatory factor analysis for applied research. New York: The Guilford Press; 2006.
25. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research; 1960.
26. Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. J Rehabil Med 2003; 35: 105–115.
27. Ryan JP. Introduction to latent trait analysis and item response theory. In: Hathaway WE, editor. Testing in the schools. New directions of testing and measurement. San Francisco: Jossey-Bass; 1983, p. 49–65.
28. Marais I, Andrich D. Formalizing dimension and response violations of local independence in the unidimensional Rasch model. J Appl Meas 2008; 9: 200–215.
29. Smith EV, Jr. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J Appl Meas 2002; 3: 205–231.
30. Tennant A, Pallant J. DIF matters: a practical approach to test if differential item functioning makes a difference. Rasch Meas Transact 2007; 20: 1082–1084.
31. Tennant A, Penta M, Tesio L, Grimby G, Thonnard JL, Slade A, et al. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. Med Care 2004; 42: I37–I48.
32. Walton DM, MacDermid JC. A brief 5-item version of the Neck Disability Index shows good psychometric properties. Health Qual Life Outcomes 2013; 11: 108.
33. Johansen JB, Andelic N, Bakke E, Holter EB, Mengshoel AM, Roe C. Measurement properties of the Norwegian version of the neck disability index in chronic neck pain. Spine (Phila Pa 1976) 2013; 38: 851–856.
34. Guzman J, Hurwitz EL, Carroll LJ, Haldeman S, Cote P, Carragee EJ, et al. A new conceptual model of neck pain: linking onset, course, and care: the Bone and Joint Decade 2000–2010 Task Force on Neck Pain and Its Associated Disorders. Spine (Phila Pa 1976) 2008; 33: S14–S23.
35. Walton DM, Carroll LJ, Kasch H, Sterling M, Verhagen AP, Macdermid JC, et al. An overview of systematic reviews on prognostic factors in neck pain: results from the International Collaboration on Neck Pain (ICON) Project. Open Orthop J 2013; 7: 494–505.
36. Miles CL, Pincus T, Carnes D, Homer KE, Taylor SJ, Bremner SA, et al. Can we identify how programmes aimed at promoting self-management in musculoskeletal pain work and who benefits? A systematic review of sub-group analysis within RCTs. Eur J Pain 2011; 15: 775 e771–e711.
37. Linacre JM. Optimizing rating scale category effectiveness. J Appl Meas 2002; 3: 85–106.
38. Chen WH, Lenderking W, Jin Y, Wyrwich KW, Gelhorn H, Revicki DA. Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. Qual Life Res 2014; 23: 485–493.
39. de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide. New York: Cambridge University Press; 2011.
40. Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, et al. A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. J Clin Epidemiol 2009; 62: 288–295.