Arthritis & Rheumatism (Arthritis Care & Research)
Vol. 61, No. 4, April 15, 2009, pp 544–551
DOI 10.1002/art.24399
© 2009, American College of Rheumatology

ORIGINAL ARTICLE

Rasch Analysis Provides New Insights Into the Measurement Properties of the Neck Disability Index

GABRIELLE VAN DER VELDE,[1] DORCAS BEATON,[2] SHEILAH HOGG-JOHNSTON,[3] ERIC HURWITZ,[4] AND ALAN TENNANT[5]

Objective. **The most widely used neck-specific measure in intervention trials is the 10-item Neck Disability Index (NDI), which is assumed to be a unidimensional interval scale, as shown by how NDI data are scored, analyzed, and interpreted. Our objective was to use modern measurement methods to test this assumption (and thereby to also test the validity of calculating summed scores and parametric statistics on NDI data) through Rasch analysis.**

Methods. **NDI data from 521 trial subjects with neck pain were fit to the Rasch model. We examined threshold ordering of NDI items, fit of data to model expectations, presence of differential item functioning, and whether or not the set of NDI items collectively measure a single construct, which is a requirement for calculating summative scores.**

Results. **There was a lack of fit of data to the Rasch model (χ^2 = 140.35, 70 df; $P < 0.001$). Five items (personal care, lifting, headaches, work, and recreation) had disordered response thresholds. Differential item functioning was detected for age and sex. The NDI items did not contribute to a single construct. Unidimensionality and interval scaling were achieved by removing 2 of the 10 items (resulting in the NDI-8), and converting NDI-8 ordinal (paper) summative scores to NDI-8 interval (Rasch-weighted) scores.**

Conclusion. **As originally proposed and conventionally used, the NDI is not a unidimensional scale, and has only ordinal scaling. This raises fundamental doubts about the practice of calculating change scores and other parametric statistics on NDI data. A revised 8-item version provides unidimensional interval-level measurement of neck pain disability.**

INTRODUCTION

The Neck Disability Index (NDI) is the most widely-used neck-specific measure in published intervention studies (1–13). It has been cited in over 251 scientific articles and translated into more than 20 languages (available at http://www.cmcc.ca/Portals/0/pdfs/rs_%20NDI_manual.pdf). The NDI was developed to assess the effect of patients' neck pain on their activities of daily living, and was based on the Oswestry Low Back Pain Disability Questionnaire, a region-specific questionnaire for patients with low back pain (14). The NDI consists of 5 items derived from the Oswestry Questionnaire (14) and 5 items identified by literature review and clinicians, with minimal input from patients. When published in 1991, the NDI was the first self-administered instrument to measure disability related to neck pain (1).

The NDI is assumed to be a unidimensional interval scale (as shown by how its response data are scored, analyzed, and interpreted) (15–17), but this assumption has not been tested. Verifying the assumption that the NDI is a unidimensional interval-level scale is important. Unidimensionality is necessary to validly summate NDI response data into a single score, and is a fundamental requirement of construct validity (18,19). Interval-level measurement is a requisite assumption of the parametric statistics (such as means, effect sizes, and minimal detectable change) that are used to analyze outcomes and compare treatment responses across groups (18,20,21). Interval-level scaling allows for the straightforward interpretation of NDI change scores where, for example, a change

Dr. van der Velde's work is supported by a Fellowship award and a Bisby award from the Canadian Institutes of Health Research. Dr. Beaton's work is supported by a New Investigator award from the Canadian Institutes of Health Research.

[1]Gabrielle van der Velde, DC, PhD: Toronto Health Economics and Technology Assessment Collaborative, University of Toronto, Toronto General Research Institute, University Health Network, Institute for Work & Health, Centre of Research Expertise in Improved Disability Outcomes, University Health Network Rehabilitation Solutions, Toronto Western Hospital, Toronto, Ontario, Canada; [2]Dorcas Beaton, PhD: Institute for Work & Health, University of Toronto, and Li Ka-Shing Knowledge Institute of St. Michael's Hospital, Toronto, Ontario, Canada; [3]Sheilah Hogg-Johnston, PhD: Institute for Work & Health, and the University of Toronto, Toronto, Ontario, Canada; [4]Eric Hurwitz, PhD: University of Hawaii at Mānoa, Honolulu; [5]Alan Tennant, PhD: University of Leeds, Leeds, UK.

Address correspondence to Gabrielle van der Velde, DC, PhD, Toronto Health Economics and Technology Assessment Collaborative, Leslie Dan Pharmacy Building, University of Toronto, 6th Floor, Room 658, 144 College Street, Toronto, Ontario, Canada M5S 3M2. E-mail: gabrielle.vandervelde@theta.utoronto.ca.

Submitted for publication May 3, 2008; accepted in revised form January 14, 2009.

in score from 25 to 20 is equivalent to a change from 10 to 5. This is a characteristic of fundamental measurement and, in order to provide this, NDI data must satisfy the basic rules for constructing an interval scale as described in the theory of additive conjoint measurement (22). In contrast, ordinal scale scores represent a rank-ordering of persons on a measured trait, and relative distances between scores are meaningless (23). Thus, if we wish to use the NDI to calculate change scores, or apply parametric statistics given appropriate distributions, we must obtain a measure of neck pain disability that is free from bias and satisfies the rules for constructing interval scale data.

Although the measurement properties of the NDI have been extensively examined using classic test theory methods (1,4,15–17,24,25), the assumption of interval-level measurement of neck pain disability has not been verified. Our objective was therefore to use modern psychometric approaches, specifically Rasch analysis, to determine whether the NDI is a unidimensional interval scale, and, if necessary, to alter the NDI using the fewest modifications possible to have it achieve interval-level scaling.

PATIENTS AND METHODS

NDI description. The NDI consists of 10 items: pain intensity, personal care, lifting, reading, headaches, concentration, work, driving, sleeping, and recreation. Each item has 6 response categories ranging from 0–5, where 0 describes the best situation and 5 describes the worst. The response categories are ordered, adjectival categories that differ descriptively across items, and respondents are asked how their neck pain affects each item. NDI scores are calculated by summing the numeric response to all items to generate a final score that ranges from 0–50. The developers of the NDI have suggested an interpretation for these scores in which a score of $\leq$4 corresponds to no disability and scores $\geq$35 correspond to complete disability (1). Minimal clinically important change and minimal detectable change have been estimated to be 3.5 NDI points and 5 NDI points, respectively (16,17).

Patients and setting. We pooled subject-level baseline data from 2 randomized trials of neck pain treatments, including age, sex, level of education, neck pain intensity, and NDI responses (26,27). Data were pooled in order to achieve an adequate sample size as recommended by Linacre (28). Patients had a primary problem of mechanical (nonspecific) neck pain and were recruited in the first trial through the membership of a Southern California health maintenance organization, having presented with neck pain for treatment (27), or in the second trial through newspaper advertisement (26). Potential subjects were excluded from the trials if they had 1) neck pain due to fracture, severe spondylarthropathy, myelopathy, vascular disease, infection, tumor, or other nonmechanic causes of neck pain; 2) progressive neurologic deficit or severe coexisting disease; or 3) pain involving third-party liability or a compensation claim.

Rasch analysis. Rasch analysis is a process of testing whether data from a scale, such as the NDI, satisfies the rules for constructing interval scale measurement. In doing so, it uses a measurement model developed by the Danish mathematician Georg Rasch (29), which provides a template for such measurement against which the pattern of observed data can be contrasted. Where the observed pattern of response does not deviate from that expected by the model, a transformation of ordinal into interval scaling is achieved. The process of Rasch analysis has been described in detail elsewhere (30,31). Briefly, the process involves testing a number of attributes of the scale and assumptions of the model. Where the items have more than 2 response options, as in the case of the NDI, an increase in response option should reflect an increase in the underlying trait. If this does not occur, the transition points (thresholds) between adjacent response options are said to be disordered. This may necessitate collapsing of categories to ensure a correct ordering of thresholds.

When observed responses are equivalent or do not greatly differ from the expected responses from the model, data are said to fit the Rasch model. A series of summary and individual item and person fit statistics are used to test the difference between observed responses and responses expected from the model (31). Generally, chi-square fit statistics are required to be nonsignificant (Bonferroni adjusted). Residual fit statistics are expected to be within a given range: $\pm$2.5 for individual items, and with a mean fit residual value close to 0.0 and an SD approaching 1.0 (usually $<$1.4) for summary statistics.

Item bias or differential item functioning can also contribute to item misfit. Differential item functioning is a form of bias in which one subgroup (e.g., women) with given levels of a latent trait responds differently to an item compared with another subgroup (e.g., men) with similar levels of that latent trait. Differential item functioning is detected by conducting an analysis of variance for each item, comparing across levels of subject characteristics and levels of latent trait. We tested differential item functioning for age (19–30 years, 31–45 years, 46–59 years, and 60–78 years) and sex (women, men).

A person separation index, which is an estimate of a scale's internal consistency, was also calculated. This index is calculated using the same formula as Cronbach's alpha (32), except that logit scale estimates for each person derived from the Rasch analysis are used, rather than their raw scores. We considered $\geq$0.70 acceptable because our focus was on group-level measurement (33).

The process of Rasch analysis also involves testing the assumption of local independence (34). Local independence assumes that a response to an item is independent of responses to other items, after controlling for the latent trait (35). Breaches of this assumption can arise because of local response dependency, in which the answer to one item would determine the answer to another, and multidimensionality. These are tested by an examination of the residuals, with the former tested by looking for residual correlations $\geq$0.03 and the latter tested by generating estimates from sets of items that represent high-positive and high-negative loadings of the first component of the residuals (36). These estimates are compared by a series of

Table 1. Summary of Rasch analyses of the Neck Disability Index*

Analysis no., description	Item fit residual, mean ± SD	Person fit residual, mean ± SD	Item-trait total chi-square		PSI	Test of unidimensionality, percent (95% CI)
			χ^2 (df)	*P*		
1. Initial Rasch analysis (n = 521)	0.57 ± 2.13	−0.22 ± 1.06	140.35 (70)	< 0.001	0.77	9.6 (7.7–11.5)
2. Headaches, lifting deleted (n = 521)	0.07 ± 1.22	−0.30 ± 0.98	62.73 (56)	0.25	0.76	4.4 (2.5–6.3)
3. Random sample 1 (n = 260)	0.26 ± 1.08	−0.27 ± 0.95	65.52 (56)	0.18	0.80	6.9 (4.3–9.6)
4. Random sample 2 (n = 260)	0.13 ± 0.83	−0.27 ± 0.95	56.42 (56)	0.46	0.76	3.5 (0.8–6.1)
5. Random sample 3 (n = 260)	0.14 ± 1.01	−0.29 ± 1.00	53.05 (56)	0.59	0.76	4.6 (2.0–7.3)
6. Random sample 4 (n = 260)	0.11 ± 1.18	−0.31 ± 1.01	70.27 (56)	0.10	0.79	6.2 (3.5–8.8)
7. Random sample 5 (n = 260)	0.18 ± 0.71	−0.27 ± 0.96	45.32 (56)	0.85	0.77	6.5 (3.9–9.2)
8. Random sample 6 (n = 260)	0.19 ± 1.01	−0.28 ± 0.97	54.00 (56)	0.55	0.76	6.9 (4.3–9.6)
9. Random sample 7 (n = 260)	0.22 ± 0.77	−0.26 ± 0.95	57.98 (56)	0.40	0.73	5.8 (3.1–8.4)
10. Random sample 8 (n = 260)	0.08 ± 0.77	−0.28 ± 0.97	53.01 (56)	0.59	0.79	5.4 (2.7–8.0)
11. Random sample 9 (n = 260)	0.10 ± 1.11	−0.29 ± 0.96	55.72 (56)	0.49	0.77	7.3 (4.7–10)
12. Random sample 10 (n = 260)	0.19 ± 0.91	−0.27 ± 0.96	45.14 (56)	0.85	0.78	4.6 (2.0–7.3)

* Random samples refer to samples of 260 patients that were randomly drawn from the pooled sample of 521 patients after the items headaches and lifting were deleted. PSI = person separation index; 95% CI = 95% confidence interval.

independent *t*-tests, and <5% should be outside of the range −1.96 to +1.96. If the lower binomial confidence interval overlaps 5%, the test is interpreted as nonsignificant and the set of items are considered to form a unidimensional scale (31). All analyses were conducted using RUMM2020 software, version 4.1 (Rumm Laboratory, Perth, Western Australia).

RESULTS

The pooled sample (n = 521) consisted of 338 women (64.88%) and 183 men (35.12%) with a mean ± SD age of 44.95 ± 11.50 years. Mean ± SD neck pain intensity in the past week was 5.17 ± 1.87 and mean ± NDI score was 13.57 ± 5.75. Some college education or less was achieved by 289 subjects (55.47%), 157 (30.13%) achieved a college education, and 75 (14.40%) achieved a professional or graduate degree. Response ranges for individual NDI item scores varied from 3–6. We did not collapse response categories selected by <1% of the subjects into the preceding adjacent category because RUMM2020 uses the conditional pairwise maximum likelihood algorithm that provides estimates in the presence of null categories (37).

Rasch analysis of data from the 10-item NDI scale demonstrated a lack of data fit to the model, with a significant item-trait total chi-square interaction (Table 1, analysis 1). The mean ± SD fit residual for items was 0.57 ± 2.13, whereas the SD should be closer to 1.0 when there is adequate fit to the model. The mean ± SD fit residual for persons was −0.22 ± 1.06, which indicated no serious misfit among subjects in the sample, on average. Two items, headaches and recreation, showed statistically significant deviation (misfit) from the model expectations (Table 2). The Bonferroni-adjusted chi-square probability values for headaches and recreation were <0.001. The fit residual values for headaches and lifting were large (4.94 and 3.42, respectively), which also suggested misfit of these items (Table 2).

The large, positive fit residual value for headaches implied a low level of discrimination. A low level of discrim-

Table 2. Individual item fit of the Neck Disability Index

Item	Location parameter	SE	Fit statistics		
			Fit residual	Chi-square	Chi-square probability*
Pain intensity	−0.85	0.06	0.59	6.28	0.51
Personal care†	1.57	0.09	−1.24	8.80	0.27
Lifting‡	0.69	0.04	3.42	16.05	0.02
Reading	0.47	0.05	−0.03	5.64	0.58
Headaches‡	−1.40	0.04	4.94	35.52	<0.00001
Concentration	0.24	0.06	−0.04	9.36	0.28
Work†	0.31	0.05	−0.84	9.12	0.24
Driving	0.17	0.06	−1.22	19.23	0.008
Sleeping	−0.55	0.05	1.44	4.60	0.71
Recreation‡	−0.66	0.05	−1.35	22.76	<0.001

* Bonferroni-adjusted probability value.
† Has disordered response categories.
‡ Misfitting (i.e., significantly deviating from Rasch model expectations and/or having a large fit residual) and has disordered response categories.

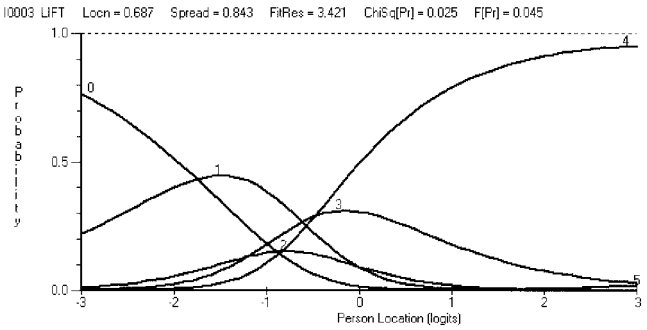

Figure 1. The category probability curves for the lifting category. LIFT = lifting; Locn = location; FitRes = fit residual; ChiSq = chi-square; Pr = probability; F = F statistic.

ination was also implied by the characteristic curve of the headaches item, in which observed responses were not as steep as the expected curve. Specifically, responses from groups with lower levels of disability were above what was expected by the model, and responses from groups with higher levels of disability were below what was expected by the model. This suggested that the headaches item captures a trait that is different than the trait captured by other items (34).

The threshold patterns for the 10 NDI items were examined on a threshold map. Distances between thresholds varied across items, which supported the use of the unrestricted (partial credit) version of the Rasch model for this analysis (38). Five items (personal care, lifting, headaches, work, and recreation) demonstrated disordered thresholds. The category probability curves for lifting, in which the third category probability curve is entirely overlapped by the other curves, are shown in Figure 1. This indicates that there is no location on the neck pain disability scale (and therefore, no level of neck pain disability) that the third response category of the lifting item was more likely to be selected than other response categories. Disordered thresholds occur when there are too many response categories or category labels are confusing. Compare, for example, the third response category for lifting, "Pain prevents me from lifting heavy weights off the floor, but I can manage if they are conveniently positioned," with the fourth response category for lifting, "Pain prevents me from lifting heavy weights, but I can manage light to medium weights if they are conveniently positioned." These response categories require the ability to clearly differentiate between heavy and light to medium weights, and it would appear that, in this sample at least, this is not the case.

The test for unidimensionality showed that positively and negatively loading sets of items provided statistically significantly different person estimates (Table 2, analysis 1). This suggested that NDI items did not contribute to a single underlying construct. Another breach of unidimensionality was the differential item functioning exhibited by the item lifting, for sex (F-ratio = 10.73, df = 1; Bonferroni-adjusted $P = 0.001$) and age (F-ratio = 5.11, df = 3; Bonferroni-adjusted $P = 0.002$). Thus, lifting did not have the same meaning in women compared with men, or in younger subjects compared with older subjects. Finally, the test of local independence identified response dependency between headaches and recreation.

The results of our initial Rasch analysis suggested that the NDI is not a unidimensional interval scale. Our objective then became to determine whether interval-level scaling could be achieved by altering the NDI with the fewest modifications possible. We proceed by removing headaches and lifting from the data to test whether adequate fit of the response data to the model could be achieved, based on the statistical results outlined above and clinical reasoning. Our clinical reasoning for removing headaches and lifting was based on a previous study in which the validity of these items was questioned because their scores did not change at treatment discharge (24). The authors surmised that headaches may not be a common symptom experienced by all neck pain patients, and therefore not sensitive to change. Similarly, patients who have not attempted to lift objects of varying weights may not know how to respond to this item.

With the removal of these 2 items, adequate fit to the model was achieved (Table 1, analysis 2). None of the remaining 8 items demonstrated statistically significant misfit or large fit residual values. None of the items comprising the NDI-8 exhibited differential item functioning to sex or age. The test of unidimensionality was nonsignificant, suggesting that the 8 remaining items contributed to a single underlying trait. There were no significant correlations between items, and only marginal local depen-

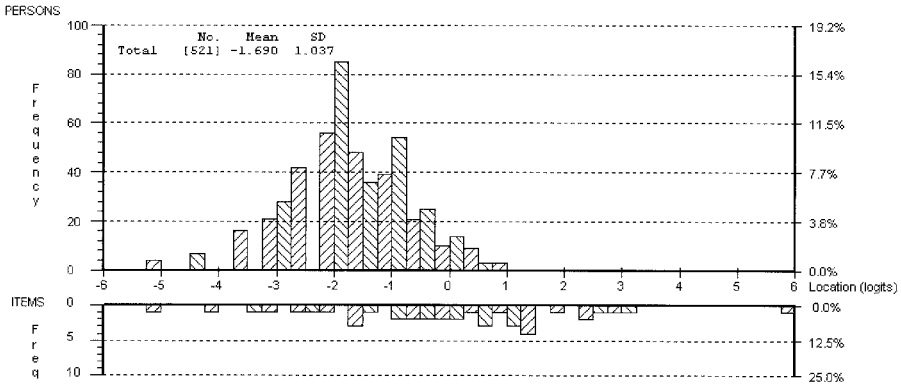

Figure 2. Distribution of subjects and 8-item Neck Disability Index item thresholds on the neck pain disability logit scale. Grouping set to interval length of 0.25, making 48 groups.

Table 3. Conversion table between Neck Disability Index 8-item (NDI-8) ordinal (summative) scores and the NDI-8 interval scores

NDI-8 ordinal score*	NDI-8 interval score†
0	0.00
1	4.14
2	7.15
3	9.36
4	11.10
5	12.53
6	13.81
7	14.93
8	15.89
9	16.82
10	17.67
11	18.45
12	19.18
13	19.88
14	20.57
15	21.23
16	21.85
17	22.47
18	23.09
19	23.70
20	24.28
21	24.86
22	25.44
23	26.02
24	26.60
25	27.19
26	27.77
27	28.35
28	28.93
29	29.54
30	30.16
31	30.82
32	31.52
33	32.17
34	32.87
35	33.60
36	34.42
37	35.42
38	36.97
39	40.72
40	50.00

* Scores obtained by summing responses to 8 items on paper.
† NDI-8 ordinal scores transformed into a 0–50 interval scale.

dency (−0.31) was identified between concentration and recreation. Personal care and work continued to demonstrate disordered thresholds.

The targeting of the NDI-8 item thresholds for subjects in our sample is shown in Figure 2. There is good coverage of thresholds over the breadth of neck pain–related disability. The mean ± SD person location value (−1.69 ± 1.04) indicates that on average, our sample is centered over the lower end of the neck pain disability scale as defined by the remaining 8 items. The person separation index (reliability) for the NDI-8 was 0.76, which is sufficient for group-level comparisons.

We verified the robustness of this NDI-8 solution by determining whether adequate fit of response data from 10 samples of 260 patients randomly drawn from our pooled sample could be achieved (Table 1, analyses 3–12). In each of these 10 samples, adequate fit of NDI-8 response data was demonstrated by nonsignificant item-trait interaction total chi-square statistics, good item and person fit to the model expectations, and nonsignificant tests of unidimensionality. The person separation index was adequate for group comparisons (range 0.73–0.80).

The NDI-8 as a simple summed ordinal scale will yield a score between 0 and 40. We transformed the NDI-8 ordinal scale into a 0–50 interval scale, using simulated data (5,000 cases) based on the item difficulties and threshold estimates of our observed response data (39). This ensured that we had a similar, equal precision of estimates for every raw score point. This provides NDI users with a straightforward conversion between NDI-8 ordinal (paper) summative scores and NDI-8 interval scores on a scale length (0–50) that NDI users are familiar with (Table 3). The lack of one-to-one concordance between ordinal and interval scores, in which ordinal scores tend to spread more widely across the end-regions of the interval scale compared with the mid-region, is illustrated in Table 3. A visualization of the spacing of ordinal scores along the NDI-8 interval scale that is required to achieve interval scaling is provided in Figure 3.

The consistencies between the original NDI ordinal scale and the NDI-8 ordinal and NDI-8 interval (Rasch-weighted) scales, estimated by an intraclass correlation coefficient (type 3,1), were 0.93 and 0.87, respectively (40). We also compared the construct validity of the original NDI scale and the interval NDI-8 scale by determining the correlation of their sum scores with neck pain intensity. For the NDI ordinal scores, the correlation between neck pain intensity and NDI ordinal scores was Spearman's ρ = 0.43 (Pearson's r = 0.42). For NDI-8 interval scores, the correlation was Spearman's ρ = 0.42 (Pearson's r = 0.45).

Thus, despite removal of 2 items and Rasch weighting of NDI scores, the internal consistency (person separation index) and construct validity of the NDI-8 interval scale was comparable with that of the original NDI ordinal scale. The NDI-8 interval scale provides the additional benefit of fulfilling criteria that must be satisfied in order to justify

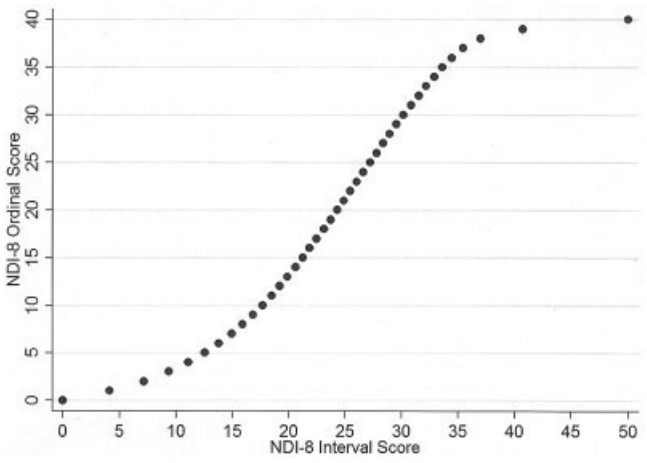

Figure 3. Plot of 8-item Neck Disability Index (NDI-8) ordinal (paper) scores versus NDI-8 interval (Rasch-weighted) scores.

the use of successive integer scores as a basis for measuring neck pain disability, and the use of mathematical functions to analyze NDI response data.

DISCUSSION

According to the Rasch model, the NDI in its original form is not a unidimensional interval-level scale, but this can be achieved by removing 2 items (headaches and lifting). This yields a valid unidimensional scale that is free of bias for age and sex. A simple summed score from the NDI-8 yields an ordinal score, but an exchange rate between the ordinal values and the interval-level measurement derived from Rasch analysis is provided in Table 3. This relationship raises fundamental doubts about the practice of reporting thresholds for meaningful change for NDI data (16,17), in that, for example, 5 raw score points on the ordinal scale relate to significantly different interval scale estimates. The difference between 5 and 10 ordinal points is 5.14 interval scale points, yet the difference between 20 and 25 ordinal points is 2.91 interval points. Thus, if use is made of the 5-point minimal clinically important difference (MCID) on the ordinal scale, MCID would be achieved much more easily by persons moving across the center of the scale than persons at the margins.

Tension between clinimetric and psychometric approaches may be created when selected items that have clinical relevance are removed. Our solution does not preclude deleted items being used for clinical management purposes, but it does preclude them being used in summative raw scores. A preponderance of evidence suggests that the NDI is a multidimensional scale. Although Hains et al concluded that the NDI is unidimensional based on exploratory and confirmatory factor analyses (25), Wlodyka-Demaille et al extracted 2 factors in a principal components analysis (41). Stratford et al found that the NDI relates to both the physical and mental component scores of the Short Form 36 health questionnaire (17). Others have concluded that the NDI is a multidimensional scale that includes symptoms (pain, headaches, concentration), impairments (personal care, lifting, reading, driving, sleeping), and disabilities (work, recreation) (42). These conflicting results might be due to limitations associated with using ordinal data in factor analysis (43).

Previous studies have specifically questioned the validity of the NDI items that we identified as misfitting (lifting, headaches) (24). Without empirical assessment, these items seem problematic. The response category labels are wordy and confusing. More than one concept is contained within the response categories (e.g., lifting ability and pain level in lifting, and headache severity and headache frequency in headaches). The items do not consist of a declarative sentence stem (namely, a single statement under consideration) to which response options represent degrees of endorsement or agreement. Rather, sentence stems are imbedded within the response options themselves.

Previous studies have also observed that response categories designed to measure the highest levels of neck pain disability are rarely endorsed by respondents (24,25,42). Personal care has been noted consistently. Washing and dressing do not appear to be often affected by neck pain to the degree described by personal care's most severe response category: "I do not get dressed; I wash with difficulty, and stay in bed." Thus, the level of disability described by NDI response categories that describe the highest levels of neck disability may be experienced rarely or by a very small proportion of neck pain patients; another possibility may be that these extreme categories are not legitimate for neck pain.

Based on the results of the best-evidence synthesis from the Bone and Joint Decade 2000–2010 Task Force on Neck Pain and Its Associated Disorders, our sample is representative of patients with mechanical (nonspecific) neck pain that are typically enrolled in intervention trials, and thus our findings should be generalizable to these patients (44). Lower levels of neck pain disability can be expected in these patients because nonspecific/mechanical neck pain tends to be used as a diagnosis of exclusion; that is, once all observable causes or lesions (e.g., cervical spine tumors, fractures, infections, inflammatory arthritides, etc.) have been reasonably ruled out. These patients have milder neck pain, and therefore they do not tend to endorse the most severe response options. This is consistent with the targeting of our current sample, which places the distribution of persons at the lower end of the operational range of the scale.

In our analysis, in which we provide an exchange rate between the NDI-8 ordinal raw score and the transformed interval scale, we chose not to correct the disordered thresholds because this would have precluded the possibility of providing a straightforward exchange between the everyday summed ordinal score and its corresponding interval score. Furthermore, collapsing the response categories would have resulted in a varying number of categories across items, which represents a considerable change from the original design of the NDI scale. Nevertheless, this disordering of response options should be further examined in other samples, preferably with more severe disability, to see if the problem is generic. If so, consideration should be given to collapsing categories at some future point, preferably in a systematic and clinically relevant way.

The NDI has not been previously evaluated with modern measurement methods. We used a methodical approach that is consistent with current recommendations (31,45). Our analysis offers new insights into the internal construct validity of the NDI, including response category functioning. We offer a modified, 8-item NDI scale that represents true (fundamental) measurement of disability related to neck pain, and because we are promoting good measurement practice, we have provided an exchange table to convert biased ordinal NDI scores to unbiased interval NDI scores. Empirical evidence from several studies suggests that outcome measurement is improved when Rasch-based scores are used (including better precision and responsiveness), and that divergent conclusions can be made when ordinal response data are treated as interval data (46–50). Nonetheless, future studies may further examine the validity of neck pain trial results when ordinal data are subjected to parametric analyses.

In conclusion, our results suggest that the NDI as it was

originally proposed and is conventionally used is not a unidimensional scale, and that it only has ordinal scaling properties. A revised 8-item version, the NDI-8, provides unidimensional interval-level measurement of disability related to neck pain.

ACKNOWLEDGMENTS

We thank Drs. Roni Evans and Gert Bronfort (Northwestern Health Sciences University, Minnesota) for generously allowing us to analyze their data for this study, as well as Dr. Peter Smith (Institute for Work & Health, Toronto) for his assistance with figures, and Mr. Mike Horton (Psychometric Laboratory for Health Services, University of Leeds, Leeds) and the Measurement Research Unit (Institute for Work & Health, Toronto) for their technical and critical advice.

AUTHOR CONTRIBUTIONS

Dr. van der Velde had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study design. Van der Velde, Beaton, Tennant.
Acquisition of data. Hurwitz.
Analysis and interpretation of data. Van der Velde, Beaton, Tennant.
Manuscript preparation. Van der Velde, Beaton, Hogg-Johnson, Hurwitz, Tennant.
Statistical analysis. Van der Velde, Beaton, Hogg-Johnson, Tennant.

REFERENCES

1. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther 1991;14:409–15.
2. Carragee EJ, Hurwitz EL, Cheng I, Carroll LJ, Nordin M, Guzman J, et al. Treatment of neck pain: injections and surgical interventions. Results of the Bone and Joint Decade 2000–2010 Task Force on neck pain and its associated disorders. Spine 2008;33:S153–69.
3. Niemisto L, Kalso E, Malmivaara A, Seitsalo S, Hurri H. Radiofrequency denervation for neck and back pain: a systematic review of randomized controlled trials. Cochrane Database Syst Rev 2003;1:CD004058.
4. Pietrobon R, Coeytaux RR, Carey TS, Richardson WJ, DeVellis RF. Standard scales for measurement of functional outcome for cervical pain or dysfunction: a systematic review. Spine 2002;27:515–22.
5. Bommireddy R, Kamat A, Smith ET, Nixon T, Pillay R, Pigott T, et al. Magnetic resonance image findings in the early postoperative period after anterior cervical discectomy. Eur Spine J 2007;16:27–31.
6. Hacker RJ. Cervical disc arthroplasty: a controlled randomized prospective study with intermediate follow-up results. J Neurosurg Spine 2005;3:424–8.
7. Jacobs WC, Anderson PG, Limbeek J, Willems PC, Pavlov P. Single or double-level anterior interbody fusion techniques for cervical degenerative disc disease. Cochrane Database Syst Rev 2004;4:CD004958.
8. Kim SH, Shin DA, Yi S, Yoon DH, Kim KN, Shin HC. Early results from posterior cervical fusion with a screw-rod system. Yonsei Med J 2007;48:440–8.
9. Latimer M, Haden N, Seeley HM, Laing RJ. Measurement of outcome in patients with cervical spondylotic myelopathy treated surgically. Br J Neurosurg 2002;16:545–9.
10. Mehren C, Suchomel P, Grochulla F, Barsa P, Sourkova P, Hradil J, et al. Heterotopic ossification in total cervical artificial disc replacement. Spine 2006;31:2802–6.
11. Peolsson A, Vavruch L, Hedlund R. Long-term randomised comparison between a carbon fibre cage and the Cloward procedure in the cervical spine. Eur Spine J 2007;16:173–8.
12. Verhagen AP, Scholten-Peeters GG, van WS, de Bie RA, Bierma-Zeinstra SM. Conservative treatments for whiplash. Cochrane Database Syst Rev 2007;2:CD003338.
13. Yoon DH, Yi S, Shin HC, Kim KN, Kim SH. Clinical and radiological results following cervical arthroplasty. Acta Neurochir (Wien) 2006;148:943–50.
14. Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry Low Back Pain Disability Questionnaire. Physiotherapy 1980;66:271–3.
15. Cleland JA, Fritz JM, Whitman JM, Palmer JA. The reliability and construct validity of the Neck Disability Index and patient-specific functional scale in patients with cervical radiculopathy. Spine 2006;31:598–602.
16. Pool JJ, Ostelo RW, Hoving JL, Bouter LM, de Vet HC. Minimal clinically important change of the Neck Disability Index and the numerical rating scale for patients with neck pain. Spine 2007;32:3047–51.
17. Stratford PW, Riddle DL, Binkley JM, Spadoni G, Westaway MD, Padfield B. Using the Neck Disability Index to make decisions concerning individual patients. Physiother Can 1999;51:107–12.
18. Spector PE. Summated rating scale construction: an introduction. Newbury Park (CA): Sage Publications; 1992.
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:S22–6.
20. DeVellis RF. A consumer's guide to finding, evaluating, and reporting on measurement instruments. Arthritis Care Res 1996;9:239–45.
21. Stevens SS. Mathematics, measurement, and psychophysics. In: Pashler H, Yantis S, Medin D, Gallistel R, Wixted J, editors. Handbook of experimental psychology. New York: Wiley; 2004.
22. Luce RD, Tukey JW. Simultaneous conjoint measurement: a new type of fundamental measurement. J Math Psychol 1964;1:1–27.
23. Embretson SE, Reise SP. Item response theory for psychologists. Mahwah (NJ): Lawrence Erlbaum; 2000.
24. Chok B, Gomez E. The reliability and application of the Neck Disability Index in physiotherapy. Physiother Singapore 2003;3:16–9.
25. Hains F, Waalen J, Mior S. Psychometric properties of the Neck Disability Index. J Manipulative Physiol Ther 1998;21:75–80.
26. 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;26:788–97.
27. Hurwitz EL, Morgenstern H, Harber P, Kominski GF, Yu F, Adams AH. A randomized trial of chiropractic manipulation and mobilization for patients with neck pain: clinical outcomes from the UCLA neck-pain study. Am J Public Health 2002;92:1634–41.
28. Linacre JM. Sample size and item calibration stability. Rasch Meas Trans 1994;7:328–31.
29. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago; 1960.
30. Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). Br J Clin Psychol 2007;46:1–18.
31. 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;57:1358–62.
32. Cronbach LJ. Coefficient α and the internal structure of tests. Psychometrika 1951;16:297–334.
33. Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, et al. Evaluating quality-of-life and health status

instruments: development of scientific review criteria. Clin Ther 1996;18:979–91.
34. Andrich D. Rasch models for measurement. Newbury Park (CA): Sage; 1988.
35. Thissen D, Steinberg L. Data analysis using item response theory. Psychol Bull 1988;104:385–95.
36. Smith EV Jr. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J Appl Meas 2002;3:205–31.
37. Luo G, Andrich D. Estimating parameters in the Rasch Model in the presence of null categories. J Appl Meas 2005;6:128–46.
38. Masters GN. A Rasch model for partial credit scoring. Psychometrika 1982;47:149–74.
39. Marais I, Luo G, Andrich D. RUMMss simulation studies program user manual. Report no. 9: maintaining invariant scales in state, national, and international level assessments. Perth, Australia: Murdoch University; 2008.
40. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;2:420–8.
41. Wlodyka-Demaille S, Poiraudeau S, Catanzariti JF, Rannou F, Fermanian J, Revel M. The ability to change of three questionnaires for neck pain. Joint Bone Spine 2004;71:317–26.
42. Hoving JL, O'Leary EF, Niere KR, Green S, Buchbinder R. Validity of the Neck Disability Index, Northwick Park Neck Pain Questionnaire, and problem elicitation technique for measuring disability associated with whiplash-associated disorders. Pain 2003;102:273–81.
43. Stout WF. A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika 1990;55:293–325.
44. Hurwitz E, Carragee EJ, van der Velde G, Carroll LJ, Nordin M, Guzman J, et al. Treatment of neck pain: noninvasive interventions. Results of the Bone and Joint Decade 2000–2010 Task Force on neck pain and its associated disorders. Spine 2008;33:S123–52.
45. Smith RM, Linacre JM, Smith EV Jr. Guidelines for manuscripts. J Appl Meas 2003;4:198–204.
46. Fitzpatrick R, Norquist JM, Jenkinson C, Reeves BC, Morris RW, Murray DW. A comparison of Rasch with Likert scoring to discriminate between patients' evaluations of total hip replacement surgery. Qual Life Res 2004;13:331–8.
47. McHorney CA, Haley SM, Ware JE Jr. Evaluation of the MOS SF-36 physical functioning scale (PF-10). II. Comparison of relative precision using Likert and Rasch scoring methods. J Clin Epidemiol 1997;50:451–61.
48. Raczek AE, Ware JE, Bjorner JB, Gandek B, Haley SM, Aaronson NK, et al. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA Project. J Clin Epidemiol 1998;51:1203–14.
49. Tennant A. Goal attainment scaling: current methodological challenges. Disabil Rehabil 2008;29:1583–8.
50. Kahler E, Rogausch A, Brunner E, Himmel W. A parametric analysis of ordinal quality-of-life data can lead to erroneous results. J Clin Epidemiol 2008;61:475–80.