J Head Trauma Rehabil

OPEN

Measuring Neurobehavioral Functioning in People With Traumatic Brain Injury: Rasch Analysis of Neurobehavioral Functioning Inventory

Karol J. Czuba, MPhty; Paula Kersten, PhD; Nicola M. Kayes, PhD; Greta A. Smith, PGDipPH; Suzanne Barker-Collo, PhD; William J. Taylor, PhD; Kathryn M. McPherson, PhD

Objective: To examine internal construct validity of the Neurobehavioral Functioning Inventory (NFI) by applying Rasch analysis. **Setting:** An outpatient rehabilitation program trial in New Zealand employing a goal-setting intervention in people with traumatic brain injury (TBI). **Participants:** One hundred eight people (mean age = 46 years; 73% male) between 6 months and 5 years post-TBI. **Design:** Rasch analysis of the NFI (Partial Credit Model). **Results:** Three NFI subscales were not unidimensional and at least 4 items in each subscale had disordered response categories. Two items showed differential item functioning by age, 1 item by educational attainment, and 2 items were found to misfit the overall construct. These items were excluded from the total score calculation. The revised scale fit the Rasch model and supported the internal construct validity of the NFI. **Conclusions:** Current scoring of the NFI subscales for people with TBI in New Zealand does not meet the requirements of the Rasch model. The revised version of NFI can improve the interpretation of scores but should be further tested with people with TBI in other settings. **Key words:** *assessment*, *measurement*, *neurobehavioral functioning*, *Rasch analysis*, *traumatic brain injury*

TRAUMATIC BRAIN INJURY (TBI) is considered to be a significant public health concern. The incidence of TBI has increased over the past decade and is estimated at 558 per 100,000 person-years in the United States,[1] with an even higher rate in New Zealand (790 per 100,000 person-years).[2] The consequences of TBI are wide-ranging with impacts on cognitive impairment being common[1] and difficult for the person with TBI and their family.[3,4] The consequent costs of these sequelae are estimated in billions of dollars annually.[5]

To improve the effectiveness and efficiency of healthcare delivery, there are increasing calls for research and clinical practice to use Patient Reported Outcome Measures (PROMs).[6,7] These measures gather patients' rather than clinicians' views on outcomes and are often more predictive of a distal outcome and quality of life than are many objective performance measures.[8] Patient Reported Outcome Measures might enhance clinicians' understanding of patients' perspectives and allow them to respond more effectively to patients' concerns. Patient Reported Outcome Measures can be used as screening tools,[9,10] as methods to help clinicians make informed decisions,[11] and as means to enhance patient-provider communication[12] and the shared decision-making process.[13,14] Furthermore, they can reduce the burden of clinical assessments and are easier to apply to large groups of patients than are structured clinical interviews.[15]

Author Affiliations: *Person Centred Research Centre, Auckland University of Technology, Auckland, New Zealand (Mr Czuba, Drs Kersten, Kayes, and McPherson, Ms Smith);Centre for Brain Research, University of Auckland, Auckland, New Zealand (Dr Barker-Collo); and Rehabilitation Teaching and Research Unit, University of Otago, Wellington, New Zealand (Dr Taylor).*

The Health Research Council of New Zealand funded the primary study, AUT University Strategic Research Investment Fund funded Karol Czuba's time on the analysis described in this paper. Kathryn M McPherson's position is part funded by the Laura Fergusson Trust Auckland.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.headtraumarehab.com).

The authors declare no conflicts of interest.

Corresponding Author: *Karol J. Czuba, MPhty, Person Centred Research Centre, Auckland University of Technology, 90 Akoranga Dr, Auckland, North Island 0627, New Zealand (kczuba@aut.ac.nz).*

DOI: 10.1097/HTR.0000000000000170

One PROM specifically developed to identify everyday problems (from symptoms to behaviors) that people with TBI experience is the Neurobehavioral Functioning Inventory (NFI). The 66-item NFI[16] was published in 1996, but a more recent version consisting of 76 items[17] published in 1999 is now widely used in both research and clinical practice. The evidence for NFI's validity and reliability rests mainly upon 2 studies by Kreutzer et al,[16,17] who reported good interrater reliability between the person with TBI and their significant other; high internal consistency with Cronbach α values exceeding 0.85 for each of the 6 subscales; and good construct validity as judged by correlation between NFI and a number of previously validated outcome measures. However, a systematic review by Cusimano et al[18] reported that the Aggression subscale did not correlate with the reference test used, suggesting inadequate construct validity for this subscale. In addition, in a recent review, Wood et al[19] argued that subsequent investigations of the factor structure of the NFI showed some variation from that originally identified, suggesting structural validity problems. In light of the widespread use of NFI, these reports are concerning and indicate a need of further refinement of the NFI subscales.

The findings suggesting issues with structural validity of the NFI[19] involved comparing results from 2 analytical approaches, namely, principal components analysis (PCA) and confirmatory factor analysis (CFA). In PCA, the variables are experientially associated, thus sample dependent.[20] In other words, the associations between variables and components are not underpinned by any prespecified theory. Conversely, CFA examines a previously hypothesized factor structure.[20] Thus, these 2 methods may yield different results. Furthermore, although both methods allow the use of nonnormally distributed data, using such data may make interpretation of the results of PCA or CFA problematic.[20]

Modern measurement theories assume that a robust scale/subscale should be unidimensional.[21] In addition, the scale should have a good spread of items along the measurement construct. These demands can be examined by the application of Rasch analysis.[21] The Rasch model can be used to examine validity, especially internal construct validity, and improve the scoring and interpretability of an established measure. When data fit the Rasch model, raw scores (ie, ordinal data) can be transformed to an interval scale.[15] The aim of this study was to apply the Rasch analysis to investigate the internal construct validity of each NFI subscale in people with TBI.

METHODS

Sample

Data was obtained from 108 participants enrolled in a randomized controlled trial exploring the efficacy of a goal-setting intervention for people with TBI. Participants were recruited in New Zealand between September 2009 and October 2011. The trial was registered with Australian New Zealand Clinical Trials Registry (ACTRN12609000433202).

Table 1 presents the demographic characteristics of the study population. Inclusion criteria included (1) a history of posttraumatic amnesia of 1 hour or more, (2) moderate disability on the Extended Glasgow Outcome Scale, (3) received compensation for at least 12 weeks, and (4) 6 to 60 months postinjury. Exclusion criteria were (1) in vegetative state and/or remaining in posttraumatic amnesia at the time of recruitment, (2) not able to demonstrate a basic level of intellectual awareness on screening, (3) unable to communicate with the researcher and/or intervention team involved in the study, and (4) unstable medical condition(s) precluding participation in rehabilitation.

TABLE 1 *Demographic characteristics of the study sample* (n = *108*)

Sex	
Females	27%
Males	73%
Age, y	
Median (IQ; Rg)	47 (56-34; 20-87)
Glasgow Coma Scale	
Median (IQ; Rg)	9 (14-5; 3-15)
Glasgow Outcome Scale—Extended	
Median (IQ; Rg)	5 (6-5; 3-6)
Posttraumatic amnesia duration, d	
Mean (SD; Rg)	20.8 (26.5; 0-120)
Time since injury, y	
Mean (SD; Rg)	1.87 (1.23; 0.41-4.76)
Qualification	
No qualification[a]	20%
High school qualification	18%
Tertiary qualification	62%
Employment prior to injury	
Full-time	70%
Part-time	8%
Unemployed	11%
Other	10%
Ethnicity[b]	
NZ European	68%
Maori	17%
Pacifica	5%
Samoan	2%
Asian	1%
Other	15%

Abbreviations: IQ, interquartile range; Rg, range; SD, standard deviation.

[a]Participants who have not completed high school and have no degree or vocational qualification.

[b]Participants identified with more than 1 ethnic group.

Data collection

The data were collected at baseline, completion of the 10-week intervention, and at 3 and 12 months postcompletion. Participants were interviewed face-to-face and also asked to complete the NFI, which was administered and scored according to the guidelines provided by the authors.[17] Participants' answers were recorded on the original response forms. The administration of the NFI in this population took between 20 minutes and 1 hour, but this information was not systematically recorded.

Instrument

As noted earlier, the NFI[17] is a 76-item self-report tool addressing a broad range of postinjury behaviors and symptoms commonly encountered by people with TBI in their daily lives. The first 6 items are labeled "critical" as they ask about symptoms that might require immediate specialist attention (eg, seizures or suicidal ideation). The remaining 70 items are organized into 6 independent subscales based on PCA and CFA carried out previously–Depression (13 items), Somatic (11 items), Memory/Attention (19 items), Communication (10 items), Aggression (9 items), and Motor (8 items). Each item is scored on a 5-point Likert-type scale as never (1), rarely (2), sometimes (3), often (4), or always (5). Scores are summed to yield total subscale scores.

Data analysis

We carried out all analyses on each of the subscales. The raw data set was prepared in SPSS 20.0 (IBM Corporation, Armonk, New York).[22] All analyses were conducted using RUMM2030 (RUMM Laboratory, Australia)[23] software to determine fit of data to the Rasch model.

The Rasch model is a mathematical framework that allows for a transformation of ordinal data to an interval scale.[24] It is based on a probabilistic relationship between responses to each item and the difference between the amount of trait possessed by the respondent ("ability") and the extent to which affirmative responses to the item measure the trait ("difficulty"). We performed a series of tests to determine whether the data met the assumptions of the Rasch model, which include a range of fit statistics. Items that did not meet these assumptions are described as misfitting. A number of factors can contribute to misfit to the Rasch model; they have been explained in detail elsewhere.[24-27] A brief outline of the analytical concepts is presented in Table 2.

Differential Item Functioning (DIF) was examined for the available grouping variables that are considered to be predictors of recovery after a TBI[32-35] and as such can affect the way the participants respond to the NFI; these include age, sex, educational attainment, and disability level (based on the Extended Glasgow Outcome Scale[36]).

We created a decision table (see Table 3), which–along with Threshold Probability Curves output from RUMM2030–guided the process of collapsing response categories in items with disordered thresholds. Achieving ordered thresholds is a fundamental concept in measurement theory, but it is also important that what we collapse makes sense.[24,37] Furthermore, we should also attempt to create a uniform frequency distribution[24] (ie, collapsing "rarely" downward with "never" rather than with "sometimes," or "often" upward with "always" rather than with "sometimes"). In the NFI, with 5 possible response categories, there are a number of collapsing strategies. The strategies in Table 3 are labeled in order of preference from first to fourth and are based on the abovementioned criteria. We continued testing possible solutions until satisfactory threshold ordering was achieved (see Table 3).

For a polytomous scale like the NFI, 1 of 2 different Rasch model parameterizations can be used. The Rating Scale Model (RSM) assumes an equal distance between thresholds across items.[38] If this assumption is not met, the current guidance is that a Partial Credit (Unrestricted) model should be used.[39] A likelihood ratio test indicated that the assumptions of the RSM were in fact not met. Consequently, we applied the Partial Credit Model, which then also showed how the assumptions of the RSM were not satisfied by revealing the disparate distances across thresholds.

For Rasch analyses, reasonably well-targeted samples of 100 are reported to have 95% confidence that the estimated item difficulty is within $\pm 1/2$ logit of its stable value.[40] Our sample of 108 participants was therefore deemed adequate for the purpose of this analysis.

RESULTS

We start by presenting the Rasch analysis results of each NFI subscales, followed by reporting the item difficulty of the Rasch-analyzed NFI.

There were no missing data in the baseline data set. The item numbers reported in this paper correspond to the item numbers assigned in the original published NFI.

Two cases were identified as potentially misfitting for the Depression subscale, 2 cases for Memory and Attention, and 1 for the Aggression subscale. Removal of these persons did not change the fit statistics. Hence, we decided to keep all cases in the data set.

Depression subscale

The initial analysis of the Depression subscale (see Table 4) showed that it was not unidimensional, and 7 of 13 items had disordered thresholds. The disordering

TABLE 2 *Brief outline of Rasch analytical concepts*

Analytical concept assessed	Test used	Acceptable values [24-26,28-30]	Strategies used to deal with misfit to the Rasch model
Person fit	Mean fit residuals (SD); range	Mean close to 0 and SD close to 1; range −2.5 to 2.5χ^2 should be nonsignificant with a Bonferroni correction	Person might have to be deleted from the data set
Item fit	Mean fit residuals (SD); range	Mean close to 0 and SD close to 1; range −2.5 to 2.5χ^2 should be nonsignificant with a Bonferroni correction	Item might have to be deleted from the subscale
Item threshold ordering[a]	Visual inspection of response thresholds for each of the items	Must show a logical progression across the trait being measured	Category responses might have to be collapsed into 1 response
Local dependence (interdependence of items)[b]	Residual item correlation matrix between all items	Correlations between the residuals <0.20 above the average residual correlation	Locally dependent items might have to be combined into testlets
Differential item functioning (DIF)[c]	ANOVA	Should be nonsignificant with a Bonferroni correction	If DIF is uniform, items might have to be combined into testlets[d] or split by person factor (eg, sex). If DIF is nonuniform, items may need to be removed.
Unidimensionality	Principal component analysis of the residuals	The 95% CI of the proportion of significant tests should include 5%	Suggests the scale is not unidimensional
Reliability index	Person separation index	Values of ≥0.70 allow for group comparisons (eg, in research trials); ≥0.85 for individual clinical use.	IF NOT applicable
Targeting of the scale to the latent trait[e]	Logit value; visual inspection of person-item distribution map	Logit value above that of the highest item on the subscale	Not applicable
Overall fit to the Rasch model	Item-trait interaction χ^2	Should be nonsignificant with a Bonferroni correction	Not applicable

Abbreviations: ANOVA, analysis of variance; CI, confidence interval; DIF, differential item functioning; SD, standard deviation.

[a]Thresholds represent points where the probability of scoring either of the 2 adjacent categories is 50%. If it is not the case, 1 would observe disordered thresholds where the individual score cannot be reliably interpreted.

[b]Local dependence occurs when a person's response to 1 item is reflected in their response to another item.

[c]DIF occurs when people from different groups (eg, males and females) with equal amounts of the underlying trait do not respond to items in a similar manner.

[d]A testlet is a bundle of items that share a common stimulus.[31]

[e]Targeting of the scale allows identification of floor and ceiling effects.

TABLE 3 *Strategy for category collapsing*[a]

Strategy	Never	Rarely	Sometimes	Often	Always
Four category					
1st	1	1	2	3	4
	1	2	3	4	4
2nd	1	2	2	3	4
	1	2	3	3	4
Three category					
3rd choice	1	1	2	3	3
4th choice	1	2	2	2	3
5th	1	1	1	2	3
	1	2	3	3	3
6th	1	1	2	2	3
	1	2	2	3	3

[a]The rationale for this process is given in the Methods section.

pattern was similar for 5 items (7, 19, 25, 43, and 65), and we collapsed responses "never" and "rarely" to create 4 categories for these items. Item 67 was also recoded into 4 categories, by collapsing responses "often" and "always." For item 59 none of the 4-category solutions worked, and ordered thresholds were achieved only when we used the predetermined sixth collapsing strategy (see Table 3).

Two sets of items were found to be locally dependent, items 7 ("feel hopeless") and 13 ("feel worthless"), and 43 ("sit with nothing to do") and 59 ("bored"), and were combined into 2 testlets. All items were free from DIF. The final solution (see Table 4) provided a good fit to the model, with item and person fit residuals within acceptable ranges, a nonsignificant item-trait interaction, good reliability, and the subscale was unidimensional. The scale was reasonably well targeted (see Figure 1, Depression) with a negative mean person location suggesting that patients had, on average, lower levels of depression than is targeted by the measure.

Somatic subscale

The initial examination of the Somatic subscale items revealed that none of the items displayed DIF, but 5 had disordered thresholds. For item 8 the third collapsing strategy resulted in best fit to the Rasch model. Items 50 and 75 were recoded by collapsing responses "never" and "rarely" into 1 category. For the remaining 2 items, 14 and 44, we collapsed responses "rarely" and "sometimes" to form 1 response category.

Evaluation of local dependence identified 4 pairs of items with correlations greater than 0.20 than mean residual correlation for this subscale. Three pairs were only marginally above the cut-off and cover somewhat different aspects of the Somatic construct. Items 8 ("stomach hurts") and 14 ("nauseous") showed residual correlation of 0.201. We combined these items into a testlet.

Finally, we evaluated the individual item fit residuals and χ^2 probabilities. Item 76 ("food doesn't taste right") showed fit residual greater than 2.5 and χ^2 Bonferroni-adjusted $P < 0.0016$. Hence, this item was deleted, resulting in an improved fit and no evidence of multidimensionality (see Table 4). The targeting of the scale was moderate (see Figure 1, Somatic).

Memory and Attention subscale

The initial analysis of the Memory and Attention subscale (see Table 4) found that the item-trait interaction was highly significant, and the scale was not unidimensional. Further evaluation identified 5 items with disordered thresholds. In 4 items (9, 21, 55, and 62), we collapsed responses "never" and "rarely" into 1 response category. For item 71, responses "often" and "always" were collapsed.

Three pairs of items were locally dependent and were combined into 3 testlets. This resulted in an improvement of summary fit statistics but did not resolve issues with local dependence, as 1 of the testlets (items 9 "forget yesterday's events" and 15 "forget if you have done things") correlated with item 51 ("misplace things"). Therefore, we combined items 9, 15, and 51 into 1 testlet, items 33 ("forget what you read") and 56 ("lose train of thought") into another, and items 55 ("concentration is poor") and 62 ("easily distracted") into the third testlet.

Further analysis identified a nonuniform DIF by age for items 39 ("lose track of time, day or date") and 64 ("forget to turn off appliances"). Deletion of these items resulted in improved fit statistics, and we did not identify DIF for any other item.

Finally, we found 2 items that were misfitting (ie, item fit residuals out of range and χ^2 probabilities below the Bonferroni adjusted threshold). Deletion of these 2 items (69 "forget to take medication" and 73 "late for appointments") improved fit of data to the model providing the final solution (see Table 4). The item and person fit residual means improved, the item-trait interaction was nonsignificant, and the unidimensionality test result was positive. The scale was reasonably well targeted (see Figure 1, Memory and Attention), with an average person indicating lower than expected levels of difficulty with Memory and Attention.

Communication subscale

The initial analysis of the Communication subscale (see Table 4) found the mean of item fit residual and standard deviation of person fit residual to be above the acceptable range. In addition, 6 items had disordered thresholds. For 3 items (28, 34, and 52) we collapsed

TABLE 4 *Rasch analysis summary statistics*

		Item fit residual[a]		Person fit residual[a]		Chi square interaction			PSI			Tests of unidimensionality 95% CI (%)	
Analysis[a]		**Mean**	**SD**	**Mean**	**SD**	**Value**	***df***	***P***	**With extremes**	**No extremes**	**α**	**Lower bound**	**Higher bound**
Depression	First[b]	0.45	1.40	−0.40	1.80	84	65	.055	0.877	0.873	0.88	6.0	14.0
	Final[c]	0.23	0.90	−0.25	1.30	57	55	.38	0.86	0.86	0.86	0.5	9.0
Somatic	First	0.36	1.24	−0.13	1.14	64	55	.19	0.76	0.75	0.8	1.0	10.0
	Final	0.33	0.94	−0.16	1.05	44	45	.48	0.73	0.72	0.77	0.5	9.0
Memory and Attention	First	0.55	1.19	−0.14	1.54	136	95	.003	0.94	0.94	0.94	13.0	22.0
	Final	0.27	0.89	−0.20	1.19	68	55	.10	0.92	0.91	0.9	2.0	10.0
Communication	First	0.66	1.13	−0.22	1.40	45	50	.65	0.85	0.85	0.84	5.1	13.0
	Final	0.44	0.91	−0.21	1.11	38	45	.73	0.84	0.84	0.84	3.0	5.0
Aggression	First	0.06	1.83	−0.26	0.95	88	45	.0001	0.84	0.84	0.86	2.0	10.0
	Final	0.03	1.12	−0.26	0.86	44	35	.14	0.82	0.82	0.86	1.0	7.0
Motor	First	0.40	1.12	−0.28	1.43	49	40	.15	0.79	0.79	0.82	2.0	10.0
	Final	0.18	0.97	−0.24	1.14	37	35	.37	0.77	0.76	0.81	1.0	7.0

Abbreviations: α, Cronbach α; PSI, Person Separation Index.
[a]Mean fit residual statistics should be close to a mean of zero with a standard deviation of 1.
[b]First refers to the analysis results of the raw ordinal data.
[c]Final refers to the analysis results of the Rasch-transformed data.

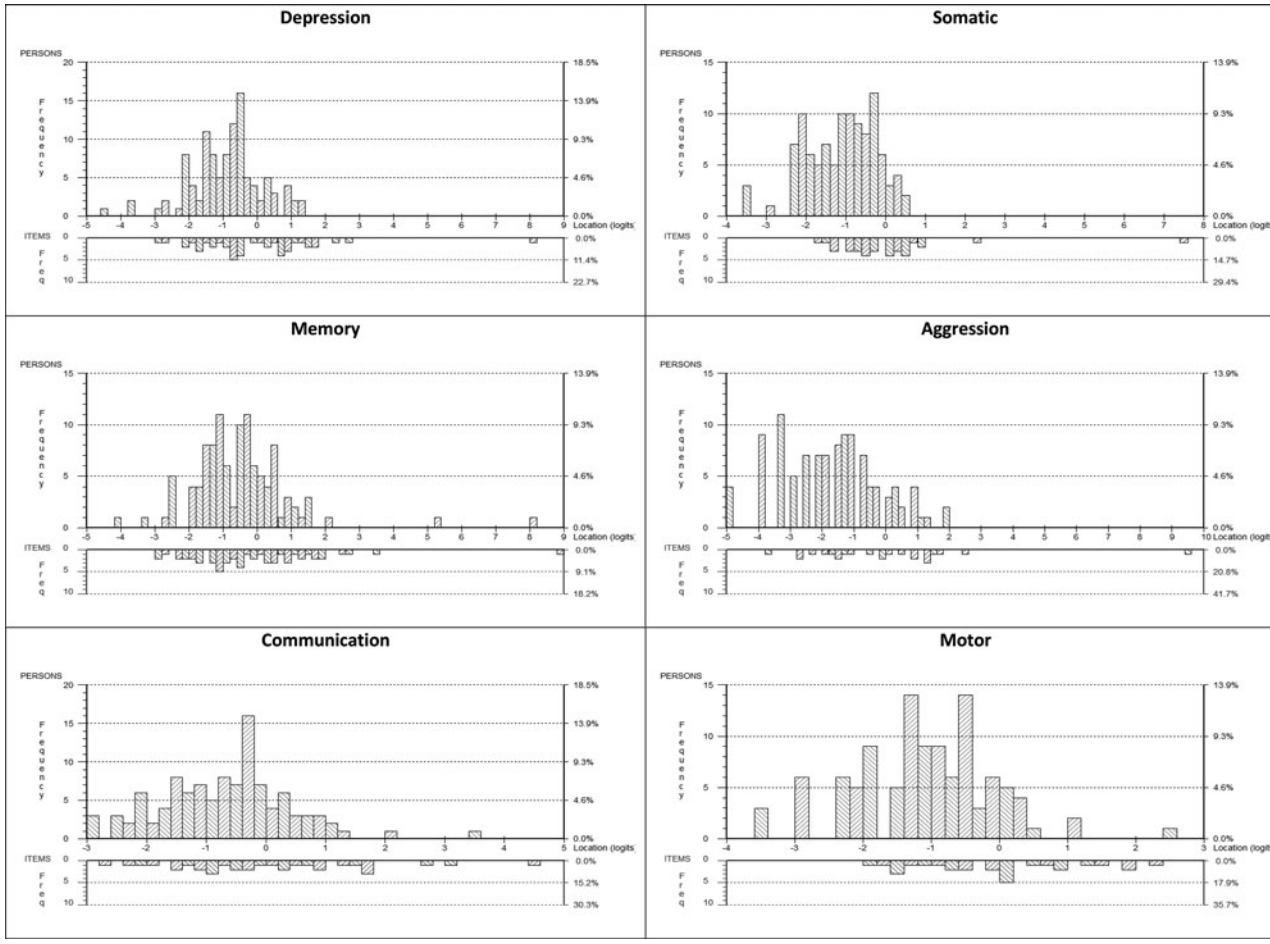

Figure 1. Person-item distribution maps for all NFI subscales.

responses "never" and "rarely" into 1 category. We found that this transformation also caused the threshold disordering for item 46 to resolve by itself. For item 16 ordering was achieved by using the third collapsing strategy (see Table 3). For item 52 only the sixth collapsing strategy resulted in threshold ordering.

Further examination identified 2 pairs of locally dependent items–items 46 ("speech doesn't make sense") with 63 ("talk too fast or slow"), and items 22 ("trouble understanding conversations") with 44 ("ringing in ears"). Both pairs were combined into testlets. No DIF was identified in this subscale. No persons were identified as misfitting. The final solution showed good fit statistics with a positive unidimensionality test result (see Table 4). The targeting of the scale was reasonably good (see Figure 1, Communication).

Aggression subscale

The initial analysis of the Aggression subscale identified some misfitting items, DIF by educational attainment and highly significant item-trait interaction. However, the subscale appeared to be unidimensional. Threshold disordering was observed in 6 items. For 5 items (17, 29, 35, 47, and 68) we collapsed responses "often" and "always" into 1 category. For item 53 the third collapsing strategy achieved threshold ordering and provided the best fit to the data.

Examination of item 53 ("threaten to hurt others") showed DIF by educational qualification. This item was deleted from the Aggression subscale.

Analysis of individual item fit residuals found a misfit for item 68 ("curse at yourself") with fit residual of 3.349 ($P < 0.002$). Deletion of this item improved item fit residual and provided the final solution with a nonsignificant item-trait interaction and a positive result from the test of unidimensionality (see Table 4). The targeting of the scale was moderate with the participants showing, on average, lower than expected levels of aggression (see Figure 1, Aggression).

Motor subscale

The initial analysis of the Motor subscale data set (see Table 4) showed some misfit to the model, indicated by the item and person fit residuals, and thresholds

disordering. First, we collapsed response categories in all items with disordered thresholds (12, 18, 42, and 48) by collapsing responses "never" and "rarely" into 1 category.

Evaluation of residual correlations showed local dependence for 1 pair of items, 36 ("drop things") and 54 ("trip over things"). We created a testlet for the 2 items. This resulted in further improvement of item and person fit residuals (see Table 4). No DIF was identified. No persons were identified as misfitting. The targeting of the scale was moderate (see Figure 1, Motor).

Item difficulty

Table 5 presents the relative difficulty of each item of the Rasch-analyzed NFI. Easy items are expected to be scored high by persons with high levels of investigated construct, whereas difficult items are expected to be scored low by persons with low levels of construct.

DISCUSSION

This investigation was undertaken to guide analysis and interpretation of NFI scores for research and clinical practice. Although the analysis identified a number of issues that affect the psychometric properties of the NFI, adjustments to scoring are proposed to fit item-responses to the Rasch model, and we suggest the use of conversion tables (see Supplemental Digital Content, available at http://links.lww.com/JHTR/A144) to improve interpretation of NFI subscale scores.

Our analyses identified that the original way responses are scored on the NFI is not always optimal with 4 or more items in each of the 6 NFI subscales having disordered response thresholds. This suggests that respondents find it difficult to distinguish between 2 adjacent response categories. If people cannot make the distinction between, for example, "sometimes" and "often," it is likely that they will use these responses inconsistently. This introduces unreliability to the measure[24] and consequently summation of scores "does not make sense."[41] When the analysis of a rating scale indicates inconsistent use of response categories (eg, disordered thresholds), the neighboring categories need to be reviewed and likely combined. As such criteria are not examined with classical test theory approaches, our analysis therefore makes a unique contribution to the assessment of the NFI response categories and to interpreting NFI scores.

One approach to the Rasch analysis is to derive a pure scale in which all problematic items (eg, with disordered thresholds, DIF, and local dependence) are deleted. However, in the case of an existing, widely used scale, it is preferable not to eliminate items, so that the original scale is retained as much as possible, while ensuring acceptable fit to the Rasch model. In our analyses of

TABLE 5 *Item difficulty*

Item difficulty	Depression		Somatic		Memory and Attention		Communication		Aggression		Motor	
Less	Item	Location	Item	Location	Item	Location	Item	Location	Item	Location	Item	Location
	37	−1.249	74	−1.034	27	−1.193	34	−0.974	29	−1.162	36 and 54[a]	−0.182
	72	−0.901	20	−0.852	57	−0.761	61	−0.782	11	−0.845	12	−0.149
	67	−0.676	44	−0.140	33 and 56[a]	−0.663	28	−0.686	47	−0.297	18	−0.079
	70	−0.611	75	−0.116	71	−0.613	16	−0.547	41	0.161	24	−0.222
	31	−0.396	32	0.071	9 and 15 and 51[a]	−0.439	10	0.127	35	0.439	30	−0.447
	7 and 13[a]	−0.180	26	0.089	55 and 62[a]	−0.158	63	0.431	17	0.678	42	0.514
	65	0.337	50	0.105	66	−0.035	52	0.627	23	1.026	48	0.566
	25	0.476	8 and 14	0.287	45	0.166	22 and 40[a]	0.695				
	43 and 59[a]	0.571	38	1.590	60	0.352	46	1.110				
	19	0.595			21	1.014						
More	49	2.034			58	2.331						

[a]A testlet made of 2 or more original NFI items. Testlet is scored by summing the scores from included items.

subscales, between 26% and 66% of items had reversed thresholds. In addition, 3 items displayed DIF, and 12 pairs of items appeared to be locally dependent. Thus, had we taken a purist approach, the subscales would have become too short for meaningful clinical use. Furthermore, although knowing a total subscale score is important, scores on individual items can indicate the specific nature of difficulty and can guide the clinical decision-making process. For example, item 64 ("forget to turn off appliances") was identified as redundant from the Rasch model perspective. However, it provides information that can be crucial to a person's safety.[24] Given its widespread use, we suggest keeping the maximum number of items to maintain the face validity and clinical utility of the NFI. Yet, when scoring patients' performance, items which have disordered thresholds need to be rescored as we have outlined above, and total raw scores need to be converted to interval-level data using a conversion table (see Supplemental Digital Content, available at http://links.lww.com/JHTR/A144).

For 4 subscales, the Person Separation Index (PSI) was more than 0.85, indicating that the subscales have the ability to differentiate well among more than 3 groups of subjects with differing amounts of the construct. These subscales are therefore suitable for use in the clinical setting.[42] For the Motor and Aggression subscales, the PSI indices were below the 0.85 threshold (0.77 and 0.82, respectively). As argued by Streiner and Norman,[42] such values indicate that the tool can only distinguish between 2 groups of patients. Thus, these subscales lack the precision needed in a clinical tool but can be used to compare groups of patients (eg, in research trials).[42] It also implies that a single total score of Motor or Aggression traits may not be a sufficient indicator of person's functioning. Yet even though the 2 subscales present a lower PSI than the other 4 NFI subscales, it is still above the acceptable threshold. Moreover, they may still yield clinically relevant information.

Scale unidimensionality is 1 of the basic principles of the Rasch model. Analysis of 3 NFI subscales (Motor, Aggression, and Somatic) showed positive results for tests of unidimensionality before any modifications were made. The unidimensionality of the Communication and Depression subscales was achieved after restoring the response threshold order. However, the Memory and Attention subscale, even after a number of modifications, was not unidimensional. The issue resolved only after dealing with local dependencies and the previously mentioned misfitting item 64 ("forget to turn off appliances"). To allow summation of the scores and produce a total subscale score, we deleted some items (see Results section) in the final analysis of the Somatic, Aggression, and Memory and Attention subscales. This is essential to obtain accurate and valid scores for individuals. From a clinical perspective, these items may provide potentially important information about the patient. For example, even though the item "forget to turn off appliances" did not fit the Rasch model, it conveys important information for a clinician. Therefore, although this item cannot be included in the total score calculation, it could be retained as part of the NFI for patient management purposes. We suggest moving such misfitting items to the end of the form to prevent their accidental inclusion into the total scores.

CONCLUSIONS

The NFI is an established and widely used measure. It addresses important aspects of neurobehavioral functioning of people with TBI. Our analysis has shown problems in the current scoring for each of the subscales in the New Zealand sample. Drawing on the findings of disordered thresholds and local dependence, we suggest that a revised scale should be further tested with people with TBI in other settings. In the meantime, we make a rescoring and conversion Excel sheet available for use to others (see Supplemental Digital Content, available at http://links.lww.com/JHTR/A144). This will ensure a valid and precise assessment of patients' functioning and conversion of NFI data to interval data, allowing for meaningful interpretation of the scores and the use of parametric statistical analyses.

REFERENCES

1. Leibson CL, Brown AW, Ransom JE, et al. Incidence of traumatic brain injury across the full disease spectrum: a population-based medical record review study. *Epidemiology*. 2011;22(6):836–844.
2. Feigin VL, Theadom A, Barker-Collo S, et al. Incidence of traumatic brain injury in New Zealand: a population-based study. *Lancet Neurol*. 2013;12(1):53–64.
3. Foster AM, Armstrong J, Buckley A, et al. Encouraging family engagement in the rehabilitation process: a rehabilitation provider's development of support strategies for family members of people with traumatic brain injury. *Disabil Rehabil*. 2012;34(22):1855–1862.
4. Wade SL, Taylor HG, Drotar D, Stancin T, Yeates KO, Minich NM. A prospective study of long-term caregiver and family adaptation following brain injury in children. *J Head Trauma Rehabil*. 2002;17(2):96–111.
5. Frost RB, Farrer TJ, Primosch M, Hedges DW. Prevalence of traumatic brain injury in the general adult population: a meta-analysis. *Neuroepidemiology*. 2013;40(3):154–159.
6. Reeve BB, Wyrwich KW, Wu AW, et al. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Qual Life Res*. 2013;22(8):1889–1905.

7. Wilde EA, Whiteneck GG, Bogner J, et al. Recommendations for the use of common outcome measures in traumatic brain injury research. *Arch Phys Med Rehabil*. 2010;91(11):1650–1660.e1617.
8. Friedly J, Akuthota V, Amtmann D, Patrick D. Why disability and rehabilitation specialists should lead the way in patient-reported outcomes. *Arch Phys Med Rehabil*. 2014;95(8):1419–1422.
9. Higginson IJ, Carr AJ. Using quality of life measures in the clinical setting. *Br Med J*. 2001;322(7297):1297–1300.
10. Stevens AM, Gwilliam B, A'Hern R, Broadley K, Hardy J. Experience in the use of the palliative care outcome scale. *Support Care Cancer*. 2005;13(12):1027–1034.
11. Jepson RG, Hewison J, Thompson A, Weller D. Patient perspectives on information and choice in cancer screening: a qualitative study in the UK. *Soc Sci Med*. 2007;65(5):890–899.
12. Happell B, Hoey W, Gaskin CJ. Community mental health nurses, caseloads, and practices: a literature review. *Int J Ment Health Nurs*. 2012;21(2):131–137.
13. Woltmann EM, Whitley R. Shared decision making in public mental health care: perspectives from consumers living with severe mental illness. *Psychiatr Rehabil J*. 2010;34(1):29–36.
14. Greenhalgh J. The applications of PROs in clinical practice: what are they, do they work, and why? *Qual Life Res*. 2009;18(1):115–123.
15. da Rocha NS, Chachamovich E, de Almeida Fleck MP, Tennant A. An introduction to Rasch analysis for psychiatric practice and research. *J Psychiatr Res*. 2013;47(2):141–148.
16. Kreutzer JS, Marwitz JH, Seel R, Devany Serio C. Validation of a neurobehavioral functioning inventory for adults with traumatic brain injury. *Arch Phys Med Rehabil*. 1996;77(2):116–124.
17. Kreutzer JS, Seel RT, Marwitz JH. *Neurobehavioral Functioning Inventory: NFI*. San Antonio, TX: Psychological Corporation; 1999.
18. Cusimano MD, Holmes SA, Sawicki C, Topolovec-Vranic J. Assessing aggression following traumatic brain injury: a systematic review of validated aggression scales. *J Head Trauma Rehabil*. 2014;29(2):172–184.
19. Wood R, Alderman N, Williams C. Assessment of neurobehavioural disability: a review of existing measures and recommendations for a comprehensive assessment tool. *Brain Inj*. 2008;22(12):905–918.
20. Tabachnick BG, Fidell LS. *Using Multivariate Statistics*. Boston, MA: Allyn and Bacon; 2001.
21. Rasch G. *Probabilistic Models for some Intelligence and Attainment Tests*. Copenhagen, Denmark: Danish Institution for Educational Research; 1960.
22. IBM SPSS. SPSS 20.0 for Windows. New York, NY: IBM Corp; 2011.
23. Andrich D, Sheridan B, Luo G. *RUMM2030 (Computer Software and Manual). Perth, Australia: RUMM Laboratory;* 2010.
24. Bond TG, Fox CM. *Applying the Rasch Model. Fundamental Measurement in the Human Sciences*. London, England: Lawrence Erlbaum Associates; 2001.
25. Andrich D. *Rasch Models for Measurement Series: Quantitative Applications in the Social Cciences No. 68*. London, England: Sage Publications; 1988.
26. Kersten P, Kayes NM. Outcome measurement and the use of Rasch analysis, a statistics-free introduction. *N Z J Physiother*. 2011;39(2):92–99.
27. 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(8):1358–1362.
28. Goodwin LD, Goodwin WL. Focus on psychometrics. Estimating construct validity. *Res Nurs Health*. 1991;14(3):235–243.
29. Tennant A, Pallant JF. Unidimensionality matters! (A tale of two Smiths?). *Rasch Measure Trans*. 2006;20(1):1048–1051.
30. Marais I, Andrich D. Effects of varying magnitude and patterns of local dependence in the unidimensional Rasch model. *J Appl Meas*. 2008;9(2):105–124.
31. Wainer H, Kiely GL. Item clusters and computerized adaptive testing: a case for testlets. *J Educ Measurement*. 1987;24(3):185–201.
32. Cifu DX, Keyser-Marcus L, Lopez E, et al. Acute predictors of successful return to work 1 year after traumatic brain injury: a multicenter analysis. *Arch Phys Med Rehabil*. 1997;78(2):125–131.
33. Keyser-Marcus LA, Bricout JC, Wehman P, et al. Acute predictors of return to employment after traumatic brain injury: a longitudinal follow-up. *Arch Phys Med Rehabil*. 2002;83(5):635–641.
34. Perel PA. Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. *BMJ*. 2008;336(7641):425–429.
35. Brown AW, Malec JF, McClelland RL, Diehl NN, Englander J, Cifu DX. Clinical elements that predict outcome after traumatic brain injury: a prospective multicenter recursive partitioning (decision-tree) analysis. *J Neurotrauma*. 2005;22(10):1040–1051.
36. Jennett B, Snoek J, Bond MR, Brooks N. Disability after severe head injury: observations on the use of the Glasgow Outcome Scale. *J Neurol Neurosurg Psychiatry*. 1981;44(4):285–293.
37. Wright BD, Linacre JM. Combining and splitting of categories. *Rasch Measur Trans*. 1992;6(3):233.
38. Andrich D. A rating formulation for ordered response categories. *Psychometrika*. 1978;43(4):561–573.
39. Masters G. A Rasch model for partial credit scoring. *Psychometrika*. 1982;47(2):149–174.
40. Linacre JM. Sample size and item calibration stability. *Rasch Measur Trans*. 1994;7(4):328.
41. Corty E. *Using and Interpreting Statistics: a Practical Text for the Health, Behavioral, and Social Sciences*. St Louis, MO: Mosby Elsevier; 2007.
42. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to their Development and use*. Oxford, England: Oxford University Press; 2008.