Published in final edited form as:
Qual Life Res. 2012 April ; 21(3): 475–486. doi:10.1007/s11136-011-9958-8.

Neuro-QOL: quality of life item banks for adults with neurological disorders: item development and calibrations based upon clinical and general population testing

Richard C. Gershon,
Northwestern University, Chicago, IL, USA

Jin Shei Lai,
Northwestern University, Chicago, IL, USA

Rita Bode,
Northwestern University, Chicago, IL, USA

Seung Choi,
Northwestern University, Chicago, IL, USA

Claudia Moy,
National Institute of Neurological Disorders and Stroke, Bethesda, MD, USA

Tom Bleck,
Rush University Medical Center, Chicago, IL, USA

Deborah Miller,
Cleveland Clinic, Cleveland, OH, USA

Amy Peterman, and
University of North Carolina at Charlotte, Charlotte, NC, USA

David Cella
Northwestern University, Chicago, IL, USA

Richard C. Gershon: gershon@northwestern.edu

Abstract

Purpose—Neuro-QOL provides a clinically relevant and psychometrically robust health-related quality of life (HRQL) assessment tool for both adults and children with common neurological disorders. We now report the psychometric results for the adult tools.

Methods—An extensive research, survey and consensus process was used to produce a list of 5 priority adult neurological conditions (stroke, multiple sclerosis, Parkinson's disease, epilepsy and ALS). We identified relevant health related quality of life (HRQL) domains through multiple methods and data sources including a comprehensive review of the literature and literature search, expert interviews and surveys and patient and caregiver focus groups. The final domain framework consisted of 17 domains of Physical, Mental and Social health. There were five phases of item development: (1) identification of 3,482 extant items, (2) item classification and selection,

Correspondence to: Richard C. Gershon, gershon@northwestern.edu.

Electronic supplementary material The online version of this article (doi:10.1007/s11136-011-9958-8) contains supplementary material, which is available to authorized users.

(3) item review and revision, (4) cognitive interviews with 63 patients to assess their understanding of individual items and (5) field testing of 432 representative items.

Participants and Procedures—Participants were drawn from the US general population and clinical settings, and included both English and Spanish speaking subjects (N = 3,246). Confirmatory factor analysis (CFA) was used to evaluate the dimensionality of unidimensional domains. Where the domain structure was previously unknown, the dataset was split and first analyzed with exploratory factor analysis and then CFA. Samejima's graded response model (GRM) was used to calculate IRT parameters. We further evaluated differential item functioning (DIF) on gender, education and age.

Results—Thirteen unidimensional calibrated item banks consisting of 297 items were developed. All of the tested item banks had high reliability and few or no locally dependent items. The range of item slopes and thresholds with good information are reported for each of the item banks. The banks can support CAT and the development of short forms.

Conclusion—The Neuro-QOL measurement system provides item banks and short forms that enable PRO measurement in neurological research, minimizes patient burden and can be used to create multiple instrument types minimizing standard error. The 17 adult measures include 13 calibrated item banks, 3 item pools available for calibration work by others, and 1 stand-alone scale (index). The Neuro-QOL instruments provide a "common metric" of representative concepts for use across patient groups in different studies.

Keywords

Outcome measures; Quality of life; Neurological disorders; Computerized adaptive testing, item banking

Introduction

Judging the efficacy of treatment in many clinical neurology settings requires input from the patient as it relates to symptoms and functional status. Often, standard clinical tools do not assess relevant information regarding day-today functioning, especially for patients with conditions characterized by chronic pain, cognitive deficiencies, fatigue or functional decline. An effective way to judge the utility of a treatment intervention is by assessing perceived changes in symptom severity. Many traditional clinical or functional measures of disease status do not provide a comprehensive representation of the full scope of impact for a given neurological disorder or its treatment, creating a need for a patient reported outcomes (PRO) measurement tool able to incorporate various aspects of a patient's functioning, specifically cognitive, emotional, physical and social aspects of well-being [1, 2].

Neuro-QOL is a National Institute of Neurological Disorders and Stroke (NINDS)–funded initiative whose purpose is to provide a clinically relevant and psychometrically robust health-related quality of life (HRQL) assessment tool for both adults and children. The specific goals of the initiative include: (1) the development of a core set of questions that address dimensions of HRQL that are universal to patients with chronic neurological diseases, (2) the development of supplemental questions that address HRQL concerns specific to particular groups of patients based on disease status and other characteristics such as age and ethnicity, and (3) the creation of a publically available, adaptable and sustainable system allowing clinical researchers access to a common item repository and computerized adaptive testing (CAT). The measures are intended to enable the facilitation of comparisons of data across clinical trials that focus on disparate neurological disorders. From 2004 to 2010, Neuro-QOL item banks, measuring common and important HRQL life domains, were

developed, calibrated, normed and validated. This paper overviews those development efforts and presents the calibration results for the Neuro-QOL adult item banks.

Patient reported outcomes measures (PROs) enable “real time” monitoring of symptoms and quality of life which in turn can be used by clinical trialists, for comparative effectiveness research, or in clinical decisionmaking, enhancing communication between patients and their physicians. As PROs are often highly customized, conventional assessment and/or cross-study comparisons are generally impossible. Instruments created from a common calibrated item bank based upon item response theory (IRT) enable easy comparison across those instruments. Such banks consist of a large collection of items (measuring a single domain) linked on a common metric. Item banks are typically used as the “source” to create computer administered tests (CAT) and short forms [3, 4].

Methods

As the primary goal of this project was to develop an HRQL measure for widespread use in neurology clinical trials and clinical research, a key first task was to identify criteria for the acceptance of HRQL measures in these communities. We then undertook an extensive research, survey and consensus process to identify target neurological conditions, resulting in the selection of 5 adult conditions (stroke, multiple sclerosis, Parkinson’s disease, epilepsy and ALS) and 2 pediatric conditions (epilepsy and muscular dystrophies–the pediatric development efforts and resulting item banks will be discussed in another article). We identified domains through multiple methods and data sources. This included a comprehensive review of the literature and literature search, expert interviews and surveys and patient and caregiver focus groups. All of the processes listed above are described in detail in the online supplement to this manuscript. A complete description of the Neuro-QOL focus group process used to assess participants’ definition of HRQL and what areas of HRQL were most impacted by their disorder and/or treatment is described in Perez et al. [2].

In total, we fully developed instruments representing 17 domains of HRQL under three broad aspects of selfreported health (Physical, Mental and Social–see Table 1) which assess concepts universally applicable across the 5 adult disorders. Included in these domains are four additional item “pools” for domains deemed important for assessment, but of lower priority than the other domains (Bowel Function, Urinary/Bladder Function, Sexual Function and End of Life Concerns). While funding to field test and calibrate these four pools was not included as part of the contract initiative, several subsequent studies are mirroring the Neuro-QOL development methodology to create validated instruments for these additional domains.

A 6 step process to develop items

Input from patients and experts, as well as review of published literature, and an evaluation of extant questionnaire items was obtained (Institutional Review Board approval was acquired from all participating sites for these interviews and all subsequent data collection activities). The deliberative process that we followed is similar to that proposed by other large-scale item banking development projects [5–8]. This process helped to classify questions for the core item pool into content-based categories that would subsequently be analyzed. In total, there were six phases of item development: (1) identification of extant items resulting in the creation of the Neuro-QOL item library, (2) item classification and selection, (3) item review and revision, (4) cognitive interviews with patients to assess their understanding of individual items and (5) field testing [6].

Selection and development

***Step* 1—**Existing instruments and items were identified by the Neuro-QOL investigators and expert consultants through literature searches and previous item banking projects [6, 9]. The Neuro-QOL item library ultimately comprised 3,482 items. This library included information on the time frame of the response requested, the exact wording of the item stem and response options, and any context (e.g., specific instructions) for the respondent to consider when answering questions.

***Step* 2—**Once the Neuro-QOL library was populated, items were assigned to the Neuro-QOL domains through an iterative, multi-step process involving domain experts. Two independent raters worked collaboratively to sort items into sub-domains, using sub-domain content driven by item review. All discrepancies in the sub-domain allocations made by the reviewers were reconciled by an additional third reviewer to ensure consistency across domains.

***Step* 3—**Once all items were assigned to a domain area, content experts systematically deleted items from individual pools based on the following criteria: apparent semantic redundancy (i.e., availability of a more appropriately worded item); inconsistency with the domain's definition; assignment to the wrong domain area; use of vague or confusing language; lack of cultural relevance; gender inappropriateness; expected difficulty for translation; or narrowness of content or excessive disease specificity. Items that were not discarded during this process underwent an extensive review, collaboratively accomplished by two domain co-chairs and several independent content experts. The majority of the items required some revision for general consistency across banks, to assure comprehensiveness in measuring the domain; to ensure clear, understandable and precise language that both experts and respondents could understand; to facilitate linguistic translation; and/or to maintain adaptability to the data collection and analysis strategies planned. Written permission was sought for use of any items not clearly in the public domain. In the couple of cases where permission was not obtained, the items were dropped from further consideration.

***Step* 4—**Findings from individual cognitive interviews and dataset analyses were provided to content groups to integrate into their decision-making. During this process, items from PROMIS [5] and AM–PAC (Activity Measure for Post-acute Care) [10] were compared with Neuro-QOL domains and items. In cases where we had a match with PROMIS, we drew items from PROMIS and then adapted the content as needed to be true to the qualitative work we did with neurology patients (focus groups and cognitive interviews). Final item pools were reviewed by patients with the target conditions (n = 63; 9 per each of our 7 target conditions) during telephone-based cognitive interviews in English and Spanish to assess the content validity of items, clarify concepts and refine language and response options. During interviews, patients reviewed each item in a one-on-one semi-structured interview focused on item comprehension and relevance. Patients and experts also identified areas for new item development and creation, to which additional items were written or revised.

***Step* 5—**Overall, the primary goal of field testing was to use the data to better understand the dimensional structure of items that specifically pertained to various domain areas of Neuro-QOL. Additionally, results were intended to inform the revision of items in the item pools and facilitate new item development prior to the first wave of testing. Prior to field testing, instruments were translated into Spanish using a rigorous process which we have utilized in other item bank development projects to maximize the similarity of responding for patients regardless of language [11].

Participants and procedures for step 5–field testing

The sampling plan facilitated obtaining item calibrations for the different domain areas, estimating profile scores for varied subgroups, confirming factor structure and conducting item and bank analyses. Most of the generic item banks were field tested on samples drawn from the US general population (Wave 1b). Targeted instruments, designed solely for use with clinical populations, were only tested in an online clinical sample (Wave 1a). A subsequent round of in-clinic testing (Wave 2) was conducted to confirm and/or improve IRT parameters and to validate the instruments in clinical settings. Data from in-clinic testing were combined with the general population data for item banks where extreme items were not endorsed with sufficient frequency in the general population sample. Sleep Disturbance was tested in both the online clinical sample as well as an in-clinic sample in order to have sufficient sample size relevant to the item content.

Online clinical testing (Wave 1a) data were collected by YouGov/Polimetrix (www.polimetrix.com). Polimetrix's standard respondent pool for an internet-based survey is taken from a predetermined panel of people who typically respond to the company's online surveys. Chosen panelists receive modest compensation (under a $10 value) for their participation. Online general population testing (Wave 1b) data was collected through Toluna (www.toluna.com), an alternate online paneling organization, offering a similar service to that of YouGov/ Polimetrix. Toluna was chosen for Wave 1b because their fee structure was more economical for this particular sample, while their recruitment methods were similar. A second round of clinical testing (Wave 2) took place at a series of academic medical centers. Wave 2 participants were recruited from Cleveland Clinic Foundation, Dartmouth-Hitchcock Medical Center, NorthShore University Health System, Northwestern University Feinberg School of Medicine, Northwestern Medical Faculty Foundation, Rehabilitation Institute of Chicago, University of Chicago, University of Puerto Rico, the University of Texas Health Science Center at San Antonio, University of Pennsylvania, Children's Memorial Hospital and the University of California at Davis.

All participants completed a socio-demographic form consisting of approximately 20 auxiliary items measuring global health perceptions, socio-demographic variables and employment status. In addition to the item banks, subjects also responded to a series of health questions about the presence and degree of perceived limitations as they related to multiple neurological conditions affecting adults including stroke, multiple sclerosis, Parkinson's disease, epilepsy and ALS.

Wave 1a clinical sample

The Wave 1a adult clinical sample included 553 respondents (see Table 2). Please refer to Table 3 for a full breakdown of all demographic variables from all three samples.

Wave 1b general population sample

The Wave 1b adult sample included 3,123 respondents (English N = 2,113; Spanish N = 1,010–see Table 4). Each participant was assigned to complete items included in one of four forms. The instruments were assigned to specific forms to both minimize subject burden and to enable factor analyses to be conducted across similar domains (e.g., both Applied Cognition instruments were administered on single form to the same subjects). The sample was used primarily for calibrating item parameters and for establishing the midpoints of the score range for each calibrated item bank, enabling comparison of item bank scores to general population benchmark values. Item bankswere divided across a series of test forms which were administered to different samples. The primary sample demographic characteristics across forms were similar to the total Wave 1b demographics.

Wave 2 clinical samples

The Wave 2 adult sample included a total of 580 respondents accrued from 12 academic medical centers. The sample data were utilized to improve the quality of the IRT analyses for the Upper Extremity Function-Fine Motor, ADL, Lower Extremity Function-Mobility, Applied Cognition–General Concerns, Applied Cognition–Executive Function and Sleep Disturbance banks, and to conduct validation testing on short-form versions of the Neuro-QOL instruments. Subjects were compensated $20 for a baseline assessment. The validation study also included the collection of proxy data (not reported here).

Analysis plan and item calibrations

The data analysis strategy closely followed Reeve et al. [7] including evaluation of unidimensionality and estimation of item parameters using IRT models. Samejima's graded response model (GRM) [12] as implemented in MULTI-LOG [13] was used for IRT-related parameter estimations for items that met the unidimensionality requirements. GRM is a polytomous IRT model which is specifically designed for use with items with ordered categories. Differential item functioning (DIF) was assessed for gender, education and age (A summary of the analysis plan is located in the electronic supplementary material to this article, http://www.springer.com/medicine/journal/11136).

Results

Summary bank analyses including the original item bank size, identification of the calibration sample used, the number of calibrated items available for future use in the creation of IRT-based instruments (CATs and short forms), the number of uncalibrated items available for future research and the overall reliability are listed in Table 5. A summary of any significant comments regarding the statistical and IRT analyses for each of the item banks is provided in the domain-specific sections below. Unless otherwise noted there was no local dependence between items and all items fit the IRT model. All IRT parameter estimations that included a general population sample were scaled to the general population mean.

Physical Health

Following Wave 1 testing of the Lower Extremity Function and Upper Extremity Function domains, 18 Lower Extremity and 24 Upper Extremity items were dropped due to extreme skew in the score distribution. The remaining items were administered to a clinical population sample in Wave 2.

The initial Fatigue item pool consisted of 20 items. Item-total correlations were all above 0.53, with most being above 0.77. All R^2 item loadings were greater than 0.70 with the exception of one item ("Enough physical strength to do the things"), which was 0.34. The same item was rejected by S-G^2 & S-X^2 ($P < 0.01$) [14–16].

The Sleep Disturbance domain was also tested in both Waves 1 and 2. Item-total correlations for the Sleep Disturbance bank ranged from 0.363 to 0.673. Exploratory factor analysis (EFA) initially suggested a four-factor model but a subsequent parallel analysis confirmed a maximum of three factors in the data [17]. Separate scales were developed for Sleep Disturbance, Restless Leg Syndrome and Parasomnia, but model fit was acceptable only for the 10 Sleep Disturbance items (two of which were dropped due to collapsed categories). All R^2 item loadings ranged from 0.387 to 0.678.

Emotional Health

In the 30-item Depression domain, item-total correlations ranged from 0.64 to 0.90. Local dependence was identified ($r = 0.16$) between two items ("I felt like crying" and "I had crying spells"). Five items were removed from the analysis due to collapsed categories and a single item was removed for a violation of local dependence.

For the analysis of the Anxiety domain, all but 11 items had more than 40% of the sample selecting the bottom category ("never"). Item-total correlations ranged from 0.56 to 0.87. All but three items had R^2 of greater than 0.50. Sixty locally dependent pairs were identified ($r = 0.153$ to $r = 0.410$). Seven of these items were removed from further analysis.

For the analysis of the Positive Affect and Well-Being domain, all items had item-total correlation between 0.60 and 0.91, with all but one exhibiting factor loadings>0.60. Two item pairs exhibited local dependence ("Lately, I felt happy about the future" with "I was able to enjoy life" and "Lately, I had good control of my thoughts" with "Lately, I had good control of my emotions"). Two of these items plus two additional items with significant misfit were removed from subsequent analysis.

The Stigma items all demonstrated item-total correlations greater than 0.50. Three items ("People are unkind to me", "People make fun of me" and "People avoid looking at me") had frequencies of less than 5 within the highest category ("always"). EFA was conducted on the total item set and three factors were identified with the first factor accounting for 68% of the variance. Thirteen items loaded onto a first factor, which dealt with the person's reaction to the illness, two items loaded on a second factor dealing with keeping the illness from others, and twelve items loaded on a third factor dealing with people's reaction to the illness. When the two items were deleted and the analysis rerun, a two-factor structure accounted for 70% of the variance, with none of the items loading greater than 0.40 on the second factor. Local dependence ($r = 0.154$) was identified between items 5 ("people were unkind") and 6 ("made fun of by other"), and $r = 0.214$ between items 25 ("others with same illness lost jobs") and 26 ("lost friends by telling them about illness"). Using a 1-factor model, fit was minimally acceptable but the items fit a bifactor model resulting in an "essentially unidimensional" bank (as described in Gibbons and McDonald) [18].

Almost all items in the Emotional and Behavioral Dyscontrol bank had item-total correlations above 0.49, with the majority being above 0.65. EFA was conducted on the total item set and yielded two factors. The first factor accounted for 60% of the variance. After dropping two items ("problems seemed unimportant" and "hard to keep up enthusiasm to get things done"), a confirmatory factor analysis (CFA) was run on the remaining 19 items. Marginal local dependence ($r = 0.162$) was encountered between the items "hard to keep up enthusiasm" and "others said I talked in a loud or excessive manner". The item "hard to keep up enthusiasm" was subsequently deleted resulting in the creation of a single unidimensional bank.

Cognitive Health

The analysis of the Applied Cognition–General Concerns domain was initially performed using data from the general population sample. Sparse data were observed for many of the extreme item response categories. Four item pairs were identified as locally dependent. Item-total correlation ranged from 0.57 to 0.85. All items had factor loadings >0.30. Seven items were suppressed due to local dependence and/or collapsed categories. A subset of 20 items was then administered to the Wave 2 clinical sample, to obtain data from subjects who would endorse the extreme categories of some of the items. Two of these items were removed after the direction of the domain was reversed.

The initial analysis of the Applied Cognitive–Executive Function domain indicated the presence of a set of items that were unrelated to the primary factor (most of these unrelated items were related to communications and accordingly we made these items available for future research as a separate "Communications Scale"). Item responses were very much skewed, with sparse data in the extreme categories. Item-total correlations ranged from 0.54 to 0.78. All items had factor loadings greater than 0.499. Thirty-six locally dependent item pairs were identified. Thirteen items were suppressed due to local item dependence and/or collapsed categories. A subset of 20 items was then further administered to the Wave 2 clinical sample in order to obtain data from subjects who would endorse the extreme categories of some of the items. Seven items were removed incrementally in four stages due to local dependence and lack of unidimensionality/scalability and the item set was reanalyzed with 13 items. All items had R^2 of greater than 0.60.

Social Health

In the 49-item Ability to Participate domain, all item-total correlations were above 0.60. The high root mean square error of approximation statistic (RMSEA–0.224) is acceptable due to the large number of items, with all R^2 item loadings greater than 0.50. Three items ("I have to limit my regular activities with friends", "I have trouble taking care of my regular personal and household responsibilities" and "I am able to work at a volunteer job outside my home") demonstrated misfit. These items and one that exhibited poor discrimination were deleted from the analysis.

The analysis of the Satisfaction with Participation domain included 51 items. All item-total correlations were above 0.50 and all R^2 item loadings were greater than 0.30. Local dependence was observed in six item pairs in two sets, with the first measuring wishing to visit friends versus the bother with having to depend on others for help, and the second set measuring bother about depending on friends versus wishing for more socializing. Three items ("I am bothered if I have to depend on my family for help", "I am bothered if I have to depend on others for help" and "I wish I could visit my friends more often") misfit the model. A total of 6 items were deleted from the analysis.

T-scores for the Neuro-QOL banks

Table 6 provides each item bank's reliability coefficients by T-score, sample T-score means and standard deviations, and distributions by percentile. A T-score distribution has a mean of 50 and standard deviation of 10. The T-score distributions are based upon the centering of the analysis. The "CT-score" distribution is based upon a clinical sample of the relevant disease; the "GPT-score" distribution is based upon the general population sample. In all cases, higher scores indicate more of that domain (for "negative" domains, such as fatigue, a higher score is worse; for "positive" domains, such as physical function, a higher score represents higher functioning). Reliability is approximated based on the conditional SE. The range of high reliability differs between instruments.

Discussion

The National Institute of Neurological Disorders and Stroke Quality of Life (Neuro-QOL) measurement system is designed to provide item banks and short forms that enable PRO measurement that is *efficient* (minimizes patient burden without compromising reliability), *flexible* (items may optionally be used interchangeably) and *precise* (minimizing standard error). We summarized the domain framework, definitions, item pool development, and sampling plan that guided the testing and calibration of the Neuro-QOL item banks. From an item library of more than 3,000 items, we developed 17 instruments including 13 unidimensional calibrated item banks that would support computerized adaptive testing and

the future development of short forms [8], 3 item pools available for calibration work by others and 1 stand-alone scale available for future research by others. Each of these instruments, the item calibrations and related statistics will be made available for research purposes and can be readily administered online through Assessment Center (www.assessmentcenter.net) [19].

In all, 432 items were tested, with 297 items becoming part of these calibrated banks based on analysis of responses from 3,246 people in the general population and/ or clinical samples. In Cella et al. [20], we derived static short form measures for each domain, and have preliminary evidence supporting the reliability and construct validity of these item banks. Numerous additional study-tailored short forms can be created from a single bank to accommodate the special needs or preferences of individual investigators. In addition, each of the Neuro-QOL item banks can be administered using a computerized adaptive test (CAT) in which the assessment is individually tailored based upon responses to previously administered items. CAT administration reduces test length dramatically without compromising measurement precision [4, 8]. CAT simulations in support of this degree of measurement efficiency have been published on similar QOL item banks [21, 22].

The Neuro-QOL item banks and short forms are available for public use to encourage researchers of neurological diseases, across multiple settings and with a range of patient populations, to provide further validation of these instruments in additional patient populations. Complete text of each item bank and preconstructed short form can be viewed at www.neuroqol.org. The Neuro-QOL instruments provide a “common currency” of represented constructs across patient groups in different studies. In strong contrast to historically disparate measures frequently used in neurological research, Neuro-QOL provides a standard PRO measurement tool that incorporates the assessment of physical, emotional, cognitive and social patient function, heightening researcher insight into patient well-being and with the potential to directly inform clinical treatment.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Lynch EB, Butt Z, Heinemann A, Victorson D, Nowinski C, Perez L, et al. A qualitative study of quality of life after stroke: The importance of social relationships. Journal of Rehabilitation Medicine. 2008; 40:518–523. [PubMed: 18758667]
2. Perez L, Huang J, Jansky L, Nowinski C, Victorson D, Peterman A, et al. Using focus groups to inform the Neuro-QOL measurement tool: Exploring patient-centered, health-related quality of life concepts across neurological conditions. Journal of Neuroscience Nursing. 2007; 39(6):342–353. [PubMed: 18186419]
3. Hambleton, RK.; Swaminathan, H.; Rogers, HJ. Fundamentals of item response theory. Newbury Park, CA: SAGE Publications; 1991.
4. Gershon R, Cella D, Dineen K, Rosenbloom S, Peterman A, Lai JS. Item response theory and health-related quality of life in cancer. Expert Review of Pharmacoeconomics & Outcomes Research. 2003; 3(6):783–791. [PubMed: 19807355]
5. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, Ader D, Fries JF, Bruce B, Rose M. The patient-reported outcomes measurement information system (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. Medical Care. 2007; 45(5 Suppl 1):S3–S11. [PubMed: 17443116]
6. DeWalt DA, Rothrock N, Yount S, Stone AA. PROMIS Cooperative Group. Evaluation of item candidates: The PROMIS qualitative item review. Medical Care. 2007; 45(5 Suppl 1):S12–S21. [PubMed: 17443114]

7. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, Thissen D, Revicki DA, Weiss DJ, Hambleton RK, Liu H, Gershon R, Reise SP, Lai JS, Cella D. Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the patientreported outcomes measurement information system (PROMIS). Medical Care. 2007; 45(5 Suppl 1):S22–S31. [PubMed: 17443115]
8. Cella D, Gershon R, Lai JS, Choi S. The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. Quality of Life Research. 2007; 16(Suppl 1): 133–141. [PubMed: 17401637]
9. Cella D, Rothrock N, Choi S, Lai JS, Yount S, Gershon R. PROMIS Overview: Development of new tools for measuring health-related quality of life and related outcomes in patients with chronic diseases. Annals of Behavioral Medicine. 2010; 39 Suppl 1–meeting abstract.
10. Haley SM, Coster WJ, Andres PL, Ludlow LH, Ni P, Bond TL, et al. Activity outcome measurement for postacute care. Medical Care. 2004; 42(1 Suppl):I49–I61. [PubMed: 14707755]
11. Eremenco SL, Cella D, Arnold BJ. A comprehensive method for the translation and cross-cultural validation of health status questionnaires. Evaluation and the Health Professions. 2005; 28(2):212–232. [PubMed: 15851774]
12. Samejima, F.; van der Liden, WJ.; Hambleton, R. The graded response model. In: Van der Linden, WJ.; Hambleton, RK., editors. Handbook of modern item response theory. New York: Springer; 1996. p. 85-100.
13. Thissen, D. MULTILOG (Version Windows (7.0)). Lincolnwood, IL: Scientific Software International; 2003.
14. Budescu DV, Cohen Y, Ben Simon A. A revised modified parallel analysis for the construction of unidimensional item pools. Applied Psychological Measurement. 1997; 21(3):233–252.
15. Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. Applied Psychological Measurement. 2000; 24:50–64.
16. Orlando M, Thissen D. Further examination of the performance of S-X^2, an item fit index for dichotomous item response theory models. Applied Psychological Measurement. 2003; 27:289–298.
17. O'Connor BP. SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. Behavior Research Methods, Instruments & Computers. 2000; 32(3):396–402.
18. Gibbons R, Hedeker D. Full-information item bifactor analysis. Psychometrika. 1992; 57(3):423–436.
19. Gershon R, Rothrock NE, Hanrahan RT, Jansky LJ, Harniss M, Riley W. The development of a clinical outcomes survey research application: Assessment Center^{SM}. Quality of Life Research. 2010; 19(5):677–685. [PubMed: 20306332]
20. Cella D, Lai JS, Nowinski C, Victorson D, Peterman A, Miller D, Bethoux F, Heinemann A, Rubin S, Cavasos J, Reder A, Sufit R, Simuni T, Holmes G, Siderowf A, Wojna V, Bode R, McKinney N, Podrabsky T, Wortman K, Choi S, Gershon R, Rothrock N, Moy C. Neuro-QOL: Brief measures of health-related quality of life for clinical research in neurology. Submitted.
21. Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D. Efficiency of static and computer adaptive short forms compared to full length measures of depressive symptoms. Quality of Life Research. 2009; 19(1):125–136. [PubMed: 19941077]
22. Choi SW, Swartz RJ. Comparison of CAT item selection criteria for polytomous items. Applied Psychological Measurement. 2009; 33(6):419–440. [PubMed: 20011456]

Table 1

Neuro-QOL domain definitions

Physical
Function/Health
Upper extremity function-fine motor, ADL (Bank)
One's ability to carry out various activities involving digital, manual and reach-related functions, ranging from fine motor to self-care (activities of daily living)
Lower extremity function-mobility (Bank)
One's ability to carry out various activities involving the trunk region and increasing degrees of bodily movement, ambulation, balance or endurance
Bowel function (Item Pool)
Functional problems related to storage and emptying, such as incontinence or constipation, urgency, leakage or discomfort
Urinary/Bladder function (Item Pool)
Functional problems related to storage and emptying, such as incontinence, urgency, leakage or discomfort
Sexual function (Item Pool)
A person's overall evaluation of, satisfaction with and quality of sexual activities, including interest, discomfort, functioning and ability to achieve orgasm
Symptoms
Fatigue (Bank)
Sensations ranging from tiredness to an overwhelming, debilitating and sustained sense of exhaustion that decreases one's capacity for physical, functional, social and mental activities
Sleep disturbance (Bank)
Perceptions of sleep quality, sleep depth, and restoration associated with sleep; perceived difficulties with getting to sleep or staying asleep; and perceptions of the adequacy of and satisfaction with sleep
Mental
Emotional health
Depression (Bank)
Experience of loss and feelings of hopelessness, negative mood (e.g., sadness, guilt), decrease in positive affect (e.g., loss of interest), information-processing deficits (e.g., problems in decision-making), negative views of the self (e.g., self-criticism, worthlessness) and negative social cognition (e.g., loneliness)
Anxiety (Bank)
Unpleasant thoughts and/or feelings related to fear (e.g., fearfulness, feelings of panic), helplessness, worry and hyperarousal (e.g., tension, nervousness, restlessness)
Stigma (Bank)
Perceptions of self and publically enacted negativity, prejudice and discrimination as a result of disease-related manifestations
Positive affect and well-being (Bank)
Aspects of a person's life that relate to a sense of well-being, life satisfaction or an overall sense of purpose and meaning
Emotional and behavioral dyscontrol (Bank)
A set of disease and/or treatment related manifestations including disinhibition, emotional lability, irritability, impatience, and impulsiveness
End of life concerns (Item Pool)
Issues and concerns that emerge at the end of one's life (including basic functioning across physical, social, emotional, cognitive and existential domains, as well as overall satisfaction with care and symptom palliation)
Cognitive health
Applied cognition-general concerns (Bank)
Perceived difficulties in everyday cognitive abilities such as memory, attention and decision-making
Applied cognition-executive function (Bank)

	Perceived difficulties in applications of mental function related to planning, organizing, calculateig, and working with memory and learning
Communication (Scale)	
	Perceived difficulties related to oral expression, language production, articulation, comprehension and organization
Social	
Ability to participate in social roles and activities (Bank)	
	Degree of involvement in one's usual social roles, activities and responsibilities, including work, family, friends and leisure
Satisfaction with social roles and activities (Bank)	
	Satisfaction with involvement in one's usual social roles, activities and responsibilities, including work, family, friends and leisure

Item Banks are available at the project website: www.neuroqol.org

Table 2

Initial clinical sample adult enrollment (Wave 1a)

Adult Banks/scales	Number of items per form	Conditions
Socio-demographic form	20	Stroke (*n* = 209)
Clinical form	82	
Stigma Bank	26	Epilepsy (*n* = 183)
Emotional and behavioral dyscontrol Bank	20	MS (*n* = 84)
Sleep disturbance Bank	20	Parkinson's (*n* = 59)
Fatigue Bank	20	ALS (*n* = 18)
		Total = 553
		(All English-Speaking)

Table 3

Sample demographics

	Wave 1a: clinical sample	Wave 1b: general population (English-speaking)	Wave 2: clinical sample
N	553	2,113	581
Age average (SD)	56.2 (12.8)	52.67 (15.5)	55.21 (14.3)
Male	53%	50%	46%
Race			
White	95%	91%	87%
Black/African American	3%	5.5%	12%
American Indian/Alaskan Native	4%	1.5%	2%
Asian	1%	3.3%	2%
Native Hawaiian/Pacific Islander	6%	1.0%	0%
Occupation			
Homemaker	11.5%	12%	8%
Unemployed	8%	8%	9%
Retired	37%	31%	30%
Disability	26.5%	10%	34%
Leave of absence	5%	>1%	1%
Full-time employed	25%	31%	21%
Part-time employed	10%	12%	10%
Full-time student	2%	3%	1%
Marital status			
Married	60%	52%	62%
Divorced	15%	14%	11%
Widowed	7%	7%	5%
Living with someone	6.5%	7%	5%
Separated	1%	3%	2%
Never married	11%	17%	16%
Income			
>$20,000	17%	18%	16%
$20–$49,000	35%	45%	35%
$50–$99,000	30.5%	31%	28%
<$100,000	14.5%	11%	21%
Education			
Some high school or less	3.5%	2%	3%
High school or equivalent	14.5%	22%	19%
Some college	40%	40%	29%
College degree	21%	24%	29%
Advanced degree	22%	11%	20%

Table 4

General population adult enrollment (Wave 1b)

Form	Bank	Items	English	Spanish		
			Complete cases	Total	Complete cases	Total
A	Ability to participate in social roles and activities	49	429	549	177	253
	Satisfaction with social roles and activities	51				
B	Lower extremity function-mobility	38	434	518	196	254
	Assistive devices	13				
	Upper extremity function-fine motor, ADL	44				
C	Positive affect and well-being	27	484	513	234	252
	Depression	30	488		240	
	Anxiety	28	476		223	
D	Applied cognition–general concerns	46	414	533	165	251
	Applied cognition–executive function	42				

Items are available at the project website

Table 5

Neuro-QOL Item Bank Statistics

Domain	Sub-domain	Scoring direction	Item n tested	Testing sample	Item n calibrated (Uncalibrated)	Alpha	CFI	RMSEA	Slope range	Threshold range
Physical Health	Upper extremity function-Fine Motor, ADL	Better function	44	GP + C	20 (24)	0.97	0.949	0.115	2.11 to 4.68	−2.51 to −0.61
	Lower extremity function (Mobility)	Better function	37	GP + C	19 (18)	0.97	0.947	0.137	2.34 to 3.89	−3.23 to 0.39
	Fatigue	Severe symptom	20	C	19 (1)	0.99	0.942	0.192		−2.0 to 2.5
	Sleep disturbance	Severe symptom	20	C	8 (12)	0.85	0.946	0.091	1.57 to 2.89	−1.21 to 1.40
Emotional Health	Depression	Severe symptom	30	GP	24 (6)	0.98	0.966	0.098	2.42 to 5.79	−0.72 to 2.14
	Anxiety	Severe symptom	28	GP	21 (7)	0.96	0.935	0.112	1.40 to 5.52	−1.05 to 2.89
	Stigma	Severe symptom		C	24 (2)	0.95 + 0.93	0.939	0.096	1.49 to 4.19	−0.34 to 3.09
	Positive affect and well–Being	Lesser Symptom	27	GP	23 (4)	0.98	0.966	0.171	2.66 to 6.61	−1.89 to 1.47
	Emotional and behavioral dyscontrol	Severe symptom	20	C	18 (2)	0.95	0.895	0.115	2.52 to 3.52	−1.20 to 2.52
Cognitive Health	Applied cognition–general concerns	Better function	45	GP + C	18 (27)	0.94	0.938	0.098	1.80 to 4.53	−3.10 to 0.16
	Applied cognition–executive function	Better Function	45	GP + C	13 (32)	0.97	0.911	0.129	2.11 to 3.68	−4.22 to 0.05
Social Health	Ability to participate in social roles and activities	Better function	45	GP	45 (0)	0.95	0.943	0.224	2.32 to 6.38	−2.28 to 0.23
	Satisfaction with social roles and activities	Better Function	45	GP	45 (0)	0.92	0.945	0.232	2.67 to 6.74	−1.88 to 0.28

Testing sample: *GP* general population (wave lb), *C* clinical (wave la), *GP + C* combined general population plus subsequent clinical (wave 2)

Table 6

Neuro-QOL item bank alpha reliability, calibration sample T-score means and standard deviations, and distributions by percentile

Item Bank	N	Center		T-scores									# Items	Mean	SD	P5	P10	P25	P50	P75	P90	P95
				10	20	30	40	50	60	70	80	90										
Anxiety	513	GP	Reliability	0.06	0.23	0.65	0.94	0.98	0.98	0.98	0.88	0.53	21	48.93	9.48	30.98	36.01	42.22	48.93	56.11	60.94	63.16
Depression	513	GP	Reliability	0.00	0.05	0.49	0.95	0.99	0.99	0.98	0.72	0.12	24	47.68	9.09	32.88	32.88	41.58	47.47	54.66	60.00	62.06
Fatigue	511	C	Reliability	0.02	0.22	0.87	0.98	0.98	0.98	0.98	0.83	0.28	19	49.76	9.93	32.88	36.45	42.82	50.01	56.95	61.55	65.64
Upper extremity function-Fine Motor, ADL	1095	GP	Reliability	0.92	0.98	0.99	0.97	0.78	0.21	0.02	0.00	0.00	20	45.12	10.85	27.28	31.05	37.42	45.10	57.00	57.00	57.00
Lower extremity function-Mobility	1046	GP	Reliability	0.77	0.97	0.98	0.98	0.96	0.74	0.15	0.01	0.00	19	47.03	9.91	30.54	33.96	39.77	46.83	54.30	62.39	62.39
Applied cognition–executive function	1109	GP	Reliability	0.90	0.96	0.97	0.96	0.89	0.56	0.13	0.02	0.00	13	47.76	9.75	31.06	35.01	41.21	47.76	54.59	60.46	60.46
Applied cognition–general concerns	1109	GP	Reliability	0.59	0.95	0.98	0.98	0.96	0.72	0.20	0.02	0.00	18	46.85	9.45	31.44	34.91	40.36	46.62	53.02	62.49	62.49
Emotional and behavioral dyscontrol	511	C	Reliability	0.05	0.28	0.78	0.95	0.97	0.97	0.97	0.95	0.84	18	49.88	9.67	34.09	38.17	43.49	49.57	56.23	62.28	64.81
Positive affect and well-being	513	GP	Reliability	0.10	0.69	0.98	0.99	0.99	0.99	0.88	0.24	0.01	23	51.28	9.82	36.03	38.78	45.69	51.80	57.67	63.17	68.32
Sleep disturbance	1087	GP	Reliability	0.09	0.30	0.60	0.81	0.88	0.90	0.89	0.85	0.72	8	49.98	9.21	35.71	38.04	43.61	49.81	56.27	61.69	65.18
Ability to participate in social roles and activities	549	GP	Reliability	0.15	0.80	0.99	0.99	0.99	0.91	0.24	0.02	0.00	45	50.43	9.56	36.10	38.62	42.79	49.04	58.58	64.91	64.91
Satisfaction with social roles and activities	549	GP	Reliability	0.06	0.59	0.98	0.99	0.99	0.88	0.12	0.00	0.00	45	50.42	9.52	36.06	38.31	42.81	49.23	58.74	63.94	63.94
Stigma	511	C	Reliability	0.01	0.06	0.31	0.84	0.98	0.99	0.98	0.95	0.69	24	49.70	9.47	35.62	35.62	41.68	50.49	56.48	61.37	64.39

GPT-scorss are based on the general population (wave lb). *CT*-scores are based upon a clinical sample (wave la). Scores derived using both *GP* and *C* samples (wave la + wave2) are centered on the *GP*. A T-score distribution has a mean of 50 and SD of 10. Reliability is approximated based on the conditional *SE*