Quality of Life Item Banks for Adults with Neurological Disorders: Item Development and Calibrations Based upon Clinical and General Population Testing

Online Supplement

Establishing Criteria for Acceptance from End Users

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. The data obtained was intended to inform the subsequent stages of the project to maximize the probability that the final HRQL instrument will meet with widespread acceptance among the likely users. To accomplish this, we set forth to obtain data on the knowledge of, attitudes toward, and use of HRQL measures and information by physicians and other professionals who participate in clinical neurology research, including clinical trials in particular. Using an adapted version of the MD-QOL, potential respondents were identified from several sources including our clinical testing sites and consultants, NINDS reviewers and grantees, the American Academy of Neurology and the American Congress of Rehabilitation Medicine. We sent a detailed letter to neurology professionals requesting their assistance and explaining our interest in developing comprehensive item banks that will be used to create generic and targeted questionnaires for clinical trials across major neurological conditions. Each respondent who expressed interest in participating was sent a web address to a password- protected, confidential online survey that contained the MD-QOL Neurology as well as demographic and qualitative questions. A total of 103 Neurology surveys were received, of which 89 had complete data. Respondents were asked to list characteristics that a HRQL questionnaire must have in order for it to be useful to the neurology clinical research community. Grouping of similar characteristics into more general categories revealed several important criteria that an HRQL measure should possess. The first criterion was satisfactory psychometric properties (38% of all responses). Respondents emphasized that the final measure should have adequate population-specific validity, reliability, responsiveness, specificity and precision and sensitivity to change. The second most frequently listed criterion was ease of administration and use (28%). The measure should be minimally burdensome for the patient and clinician, able to be completed by people at various levels of disability and literacy, include alternative modes of administration (e.g., phone, computer), use unambiguous questions and be brief (less than 10-15 minutes). The third criterion related to the content of the measure (12%) and the need to include the diversity of symptoms and HRQL domains impacted by neurological disorders. Finally, the measure should be clinically relevant and have direct application to patient care and applied outcomes (11%). The remaining responses addressed the need for the measure to be objective, have good normative data and be patient-centered.

Selection of Neurological Disorders

We first conducted interviews with 44 experts in neurological disorders and/or HRQL that asked for their opinion on: which neurological conditions to focus on and why; important HRQL domains for neurological disorders and what criteria an HRQL measure for neurological disorders should have.

Following the interviews, a consensus meeting consisting of members of the contract team, 13 independent researchers not associated with the contract, and representatives from NINDS was held for the purpose of selecting 5 adult neurological disorders to be the focus of subsequent HRQL measure development activities. Prior to selecting the target diseases, the consensus panel agreed to use the following criteria when making their decision: 1) Prevalence of the disease/disorder; 2) Magnitude of the disease’s impact on the individual; 3) The existence of promising current or new treatments on the horizon; 4) Multiple domains affected; 5) Chronic nature of the disease/possibility of seeing HRQL change.

Identifying Health Related Quality of Life Domains

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.

***Literature Review***. Initially, we identified conditions through completing an extensive Medline literature review of 24 neurological conditions using key words such as HRQL, neurological disorders, measurement issues, as well as disease-specific characteristics, from 1996 to the present. This literature review summarized major neurological disorders and their concomitant impact on HRQL.

***Expert Interviews****.* We completed two waves of expert interviews and an online request for information to identify the most important HRQL domains affected by neurological conditions. In our first wave of interviews (n=44), among other tasks, we asked respondents to identify domains or areas of QOL that are affected by neurological disorders and their treatments. Since these interviews occurred before we had selected our seven Neuro-QOL conditions (Task II), there was a low prevalence of expertise in certain conditions, such as pediatric epilepsy, ALS or muscular dystrophy. Experts were informed that their responses could include important symptoms (e.g., pain), areas of function (e.g., mobility), or anything else that was deemed important to consider when thinking of the QOL of people with neurological disorders. Experts were first asked to list all the domains they believed would be important to cover in a QOL questionnaire that could be given to patients with any neurological disorder (i.e., general neurological patients). After that, they were asked to list domains that might be important in one of the disorders they named previously, but that wasn’t necessarily common to all disorders.

We also conducted an online request for information from experts in neurology (n=89) and asked them to think about their patients who participate in clinical research and list the five neurological disorders for which they think it is most important to measure HRQL. For each disorder they were asked to identify the three areas of well-being or function (HRQL Domain) that they believe should be included in a HRQL questionnaire for that disease. These interviews confirmed domains that had been identified from the literature review and also revealed several new areas.

Finally, we completed a second wave of expert interviews (n=63) and asked them to think about all of the different ways that their subspecialty disease(s) in particular could affect patients’ quality of life and to list all domains they believe would be important to provide item coverage in a HRQL questionnaire. These interviews provided greater depth and elaboration of content for given domains.

***Patient and Caregiver Focus Groups***(N=83). We conducted 11 focus groups with patients and caregivers (seven with patients (n=64); four with caregivers (n=19)) to assess the impact of neurological conditions on quality of life domains. We began with broad questions, allowing participants to free-list responses on their definition of quality of life. We then progressed to questions regarding specific quality of life domains, such as physical function, emotional function, social aspects, and treatment effects that have been shown to be relevant in the literature. Focus groups with caregivers of Alzheimer’s Disease, stroke, and pediatric epilepsy patients were conducted as these patients may be unable to reliably report their subjective perceptions of HRQL due to cognitive impairment or age. For a more extensive review of focus group findings, please see Perez et al. [1]

***Keyword Literature Search***. Because new domains arose from these different sources, we also conducted a comprehensive keyword literature search (from 1996 to present) via OVID search engine using previous and newly identified domains and Neuro-QOL diseases to best estimate the number of published studies in a given area. We felt that these totals, albeit approximate, would provide us with an overall quantification of how important certain domains were within different neurological conditions.

***Development of a Preliminary Domain Framework*.** As a means of organizing the multiple domains that arose from the aforementioned data sources, we constructed a preliminary domain framework to guide in subsequent decision making. Our next task would be to identify which domains should be selected for generic item bank and targeted scale development.

***Selection of Generic Domains: Overall Method.*** First, we identified disease chairs from the Neuro-QOL Executive Committee and co-investigator panel. Each chair was assigned a disease from the seven Neuro-QOL conditions, which included stroke, adult epilepsy, ALS, Parkinson’s disease, multiple sclerosis, muscular dystrophy (pediatric), and pediatric epilepsy. Each chair was charged with reviewing the aforementioned data sources and extracting the most relevant domains for item bank consideration. Methods undertaken were comprehensive and iterative and involved ongoing feedback to and from the NINDS for guidance and consultation.

Each source of data was analyzed using largely qualitative approaches. Although each disease chair allowed some degree of individual judgment in making domain selection decisions, this process primarily entailed identifying and coding content derived from context specific units of meaning. These codes were converted into percentages, which were calculated as the number of times a particular theme or code was applied over the total number of all codes applied from each data source. For example, using this extraction approach it was possible for us to understand how frequently physical function was mentioned in ALS, within the context of all other domains that were mentioned for ALS. This permitted us to better understand the occurrence (and by association, importance) of certain domains either across all conditions or as a unique aspect of one disease. Constant comparison to the literature and other sources of informant data was applied and enhanced the data collection process in an iterative manner.

Within each disease, domain percentages were calculated and recorded on a chart that was populated by information obtained from five sources as mentioned above: Expert Interview I, Expert Online Request for Information, Expert Interview II, Focus Group and Keyword Literature Search. For the two expert interviews and expert survey, we made a decision to only include the open-ended, spontaneously generated expert responses (vs. information experts suggested only after being asked to elaborate on a specific domain we provided them). We felt that by only using the domain information that was a part of their “brainstorming” process, we would capture the most important areas without inadvertently biasing their thought process with domains we felt were important or had identified from other sources. If a domain met a designated cutoff criterion across all five data sources, it received a sum of “5”; if it met the cutoff across four data sources, it received a sum of “4”, and so on. These 0-5 counts were then compared across diseases. If a domain was counted as ≥3 on at least 50% of the diseases (e.g., 4/7 diseases) it was considered to be a generic concept. Targeted domains were those that summed ≥2 in at least one domain, but were not necessarily prevalent across the majority of diseases. In the event that certain disease specific domains “tied” either within or between conditions, we consulted our expert panel for their input. After reviewing the findings of this comprehensive identification and selection process, the generic domains that were chosen were: Physical, Social, Emotional and Cognitive Function.

***Selection of Generic Sub-Domains: Overall Method.*** Next, we identified domain co-chairs from the Neuro-QOL Executive Committee and co-investigator panel. Each co-chair was assigned a domain from the four generic domains previously selected and one pair was assigned the targeted domains. Each dyad was charged with reviewing the aforementioned data sources and extracting the most relevant sub-domains for item bank consideration (the dyad responsible for targeted scale decisions selected which scales should be developed and tested vs. only developed). Similar to the previous process, methods were extensive and involved continuous guidance and input from the NINDS. Again, data were analyzed using largely qualitative approaches, which are described below.

***Expert Interview II Domain Elaborations****:* During these interviews, experts (7 per Neuro-QOL condition) were asked to elaborate upon several identified domains that might be affected by the neurological condition of their sub-specialty. All responses were exhaustively coded and grouped according to their disease and domain by two outcomes researchers familiar with neurological disorders and qualitative methodologies.

Next, percentages were calculated for the total number of times a particular code was applied within a domain. This was done to crudely estimate which codes might carry additional importance for a particular domain within a disease based on how often they were discussed among experts. We anticipate using this information to appropriately portion certain types of items within a given scale.

The total number of applied codes was tallied both across and within conditions. The number of applied codes across conditions was used to determine which diseases shared similar codes relative to one another as well as which codes were unique to a particular disorder. Decisions regarding generic sub-domains were based on issues that cut across the majority of diseases. The number of applied codes within conditions was used to estimate roughly how many different issues were relevant for each disease. Decisions regarding targeted scales were based on which diseases contained the greatest number of issues, relative to other diseases.

When reviewing this data to make generic sub-domain decisions, we referred to the total number of codes across diseases as a rough indicator to determine which sub-domains are highly predominant across the majority of Neuro-QOL diseases. When applicable, we gave greater importance to sub-domains that cut across at least 5-7 diseases. For example, if bowel and bladder were to be selected as a generic sub-domain for physical function, we would note that incontinence and constipation tend to be most prevalent across conditions compared to other types of bowel and bladder issues that were coded.

***Focus Group Findings***. Focus group participants were also facilitated through domain-elaboration discussions based on pre-identified domains as well as newly elicited areas in which their respective neurological condition affects their quality of life. Responses underwent extensive open, axial and selective coding procedures by experts in qualitative methodologies. Percentages were calculated as the number of times a particular theme or code was applied over the number of all codes applied during for the focus group discussion.

When reviewing focus group data to make generic sub-domain decisions, we again referred to the frequencies and percentages of the total number of codes applied to domains across diseases. Because of the variability of responses, we simply used these percentages to provide confirmatory evidence to the expert interview data (e.g., were there any responses or not?). In some cases where focus group data were present regarding a particular sub-domain/condition association but expert interview data were not, we considered this as an area worth subsequent follow-up with other co-investigators and experts.

***Selection of Targeted Domains: Overall Method*.** Based on feedback from experts, as well as considering the complexity of issues surrounding these conditions, we decided to develop and field test select scales per condition, and also develop (but not field test) additional disease specific scales as indicated by the unique circumstances of each condition. To determine which scales would be field tested (vs. only developed), we summarized and examined data from three different sources: 1) Domain elaborations from our second round of expert interviews (N=63), findings from our focus groups (N= 11 groups), and expert panel ratings of their preferences for disease targeted scales (N= 25).

Following investigator meetings where generic domains were identified, we set forth to make preliminary decisions regarding which targeted scales should be developed, and for which disease(s). The results of these meetings led to the identification of a select number of candidate domains, which were presented to our panel of disease specific experts. Because the targeted domains presented to experts varied by disease (e.g., adult epilepsy experts were asked to rank fatigue, pain, bowel and bladder and stigma, while Parkinson’s experts were asked to rank sleep, sexual function and personality/behavioral changes) it was not possible to rank each using the same denominator, but rather to examine each disease group individually. Using these expert rankings, along with Focus Group frequencies and the total number of coded targeted domain issues within each disease we identified our candidate targeted scales to develop and field test per disease, as well as additional targeted scales for development purposes only.

When reviewing this data to make targeted scale decisions, we referred to the total number of codes by disease as a rough indicator to determine which diseases are comparatively more affected by certain issues in a given domain. When applicable, we gave greater importance to domain-condition relationships when there was an approximate and sizeable difference between total codes among conditions.

When reviewing focus group data to make targeted scale decisions, we again referred to the frequencies and percentages of the total number of codes applied to domains across diseases. Because of the variability of responses, we simply used these percentages to provide confirmatory evidence to the expert interview data (e.g., where there any responses or not?). In some cases where focus group data were present regarding a particular domain-condition association but expert interview data were not, we considered this as an area worth subsequent follow-up with other co-investigators and experts. Following this comprehensive review and selection process, the targeted scales selected for development and field testing were developed.

***Neuro-QOL “Develop but not Test” Item Pools*.** Based on findings from patient focus groups and expert interviews early in the contract period, it was determined that certain areas of functioning and quality of life were important to measure, yet were relatively lower in priority compared to other areas for which the Neuro-QOL system was developing and testing items banks. Because the scope and limitations of the contract resources and time period precluded the study team from including all possibly important domain areas, a decision was made to “develop but not test” item pools for the following areas: bowel function, bladder function, sexual function and end of life concerns. The first three domain areas were relevant to multiple diseases however the last domain was selected primarily for ALS. From the time this decision was made, item banks and scales of the same constructs have been developed and tested in related conditions or studies. For example, Bowel and Bladder/Urinary item banks have been developed and tested for men diagnosed with prostate cancer with funding from a grant from the American Cancer Society (Primary Contact: David Cella), of which Dr. Cella has agreed to make available to neurology researchers who wish to test and further validate in neurology samples. In addition, Neuro-QOL investigators (Victorson, Cella, Heinemann) have assisted in the creation of additional Bowel and Bladder item banks through Neuro-QOL extension studies, SCI-QOL and TBI-QOL (Primary Contact: David Tulsky). These banks were created using Neuro-QOL methodologies and have been administered along several Neuro-QOL item banks with large clinical samples of individuals with SCI and TBI. These are also being made available for subsequent testing and use with different neurological disorders. The PROMIS Cancer Supplement (Primary Contact: Kevin Weinfurt) has developed an extensive sexual functioning items, as well as the SCI-QOL/TBI-QOL work, and therefore both are being suggested for further consideration and use in neurology research. Finally, Neuro-QOL investigators (Victorson, Peterman) have worked in collaboration with Dr. Noelle Carlozzi’s NINDS-funded Huntington’s Disease Quality of Life study (HD-HRQL) and have helped integrate Neuro-QOL ALS focus group information into her End of Life Concerns item bank so that it can be sufficiently useful to both ALS and HD populations. Dr. Carlozzi has given permission for these items to be used in subsequent neurology research studies as a part of Neuro-QOL efforts.

Detailing Components of the Statistical Analysis Plan

In addition to descriptive statistics and item-total correlations, factor analytic approaches were used to evaluate dimensionality of items of each domain. Confirmatory factor analysis (CFA) that accounts for categorical (ordinal) data was often run without a prior exploratory factor analysis (EFA) because we designed items to fit unidimensional domains. When EFA was used, samples were randomly divided into two datasets (one for EFA and the other for subsequent CFA). Since a one-factor model is typically statistically reject able (chi-square statistic) when a large sample size is used, the focus was on practical fit indices such as the comparative fit index (criterion: CFI >0.90), Root Mean Square Error of Approximation statistics (RMSEA--criterion: < 0.10), factor loadings (criterion: >0.3), and average absolute residual correlations (criterion: <0.15). When a clinically or theoretical meaningful model was available, a bi-factor model. [2,3] was implemented to better describe the domain structure. A bi-factor model allows each item to load on a general factor and one local factor. The (squared) loadings indicate the proportion of variance of the item that is accounted for by the two factors. Those variance components are independent and measure two different variables using independent variance components. Thus, the local factors are orthogonal to each other and the general factor, and are defined by item content facets. Fitting a bi-factor model to the data allowed for the evaluation of the degree to which using an item set to scale individuals on a common factor is distorted by the presence of secondary local factors.

When Exploratory Factor Analysis (EFA) suggested a large number of factors an additional parallel analysis was conducted. In parallel analysis the size of eigenvalues expected by chance alone is compared with observed eigenvalues to help determine the maximum number of underlying dimensions.

We examined item misfit using the S-X2 and S-G2 statistics based on the observed and expected responses for each summed score. These summed score based fit indices are analogous to a Pearson’s 2 and a likelihood ratio G2 statistics, respectively. They are known to display moderate power to detect misfit with Type I error rates close to the expected nominal levels. [4-6]

Items were also examined for differential item functioning (DIF). DIF occurs when probabilities of responding in different categories differ by population for the same underlying level of the attribute. DIF analysis was conducted for gender, education and age where sufficient sample size existed. Future studies should also establish whether DIF exists for neurological disorders. Items can be evaluated for DIF by contrasting the IRT difficulty or location (*bi*) and slope (*ai*) parameters between two groups, which in this context related to the major demographic groups represented in the Neuro-QOL sample. IRT-based hierarchical ordinal logistic regression (OLR) approach as implemented in LORDIF[7] was used for evaluation of DIF. In this approach a series of logistic models predicting the probability of item response were run and compared. The independent variables in Model 1 were the trait estimate group (e.g., raw scale score) and the interaction between group and trait. Model 2 included main effects of trait and group, and Model 3 included only the trait estimate. Non-uniform DIF was detected if there was a statistically significant difference in the likelihood for Model 1 and Model 2, and uniform DIF was evident if there was a significant difference in the likelihoods for Models 2 and 3.

Following initial analysis, the results were discussed and decisions were made regarding each item. Typically, a first wave of item “cuts” was made; whereby the most problematic items were eliminated and the reduced-length item pools were subjected to follow up analyses to help arrive at decisions regarding each item. Through this process of iterative analysis and discussion with content (domain) experts, item-by-item level decisions were made determining whether an individual item should be: (1) calibrated and included in the bank, (2) not calibrated ("uncalibrated") but retained for possible future calibration (e.g., high quality items which may have clinical value in describing each domain, but otherwise did not "fit" the single factor model, exhibited local dependence, or had responses concentrated in only a few of the available response options), or (3) excluded from further consideration (e.g. outside of concept; problematic item wording).

References

1. Perez, L., Huang, J., Jansky, L., Nowinski, C., Victorson, D., Peterman, A., & Cella, D. (2007). 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, 39(6), 342-353.

2. Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. Quality of Life Research, 16 Suppl 1, 19-31.

3. Lai, J. S., Butt, Z., Wagner, L., Sweet, J. J., Beaumont, J. L., Vardy, J., Jacobsen, P. B., Jacobs, S. R., Shapiro, P. J., & Cella, D. (2009). Evaluating the dimensionality of perceived cognitive function. Journal of Pain and Symptom Management, 37(6), 982-995.

4. Budescu, D. V., Cohen, Y., & Ben Simon, A. (1997). A Revised Modified Parallel Analysis for the Construction of Unidimensional Item Pools. Applied Psychological Measurement, 21(3), 233-252.

5. Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. Applied Psychological Measurement, 24, 50-64.

6. Orlando, M., & Thissen, D. (2003). Further examination of the performance of S-X 2 , an item fit index for dichotomous item response theory models. Applied Psychological Measurement, 27, 289-298.

7. Choi, S. W., Gibbons, L. E., & Crane, P. K. (2009). Development of freeware for an iterative hybrid ordinal logistic regression/IRT DIF: A Monte Carlo simulation approach for determining cutoff values*.* Paper presented at the Annual Meeting of National Council on Measurement in Education, San Diego, CA.