Health and Quality of Life Outcomes

Research

Classical test theory versus Rasch analysis for quality of life questionnaire reduction

Luis Prieto*[1], Jordi Alonso[2] and Rosa Lamarca[2]

Address: [1]Health Outcomes Research Unit. Eli Lilly and Company, Madrid, Spain and [2]Health Services Research Unit. Institut Municipal d'Investigació Mèdica (IMIM). C/ Dr. Aiguader, 80; 08003 Barcelona, Spain

Email: Luis Prieto* - prieto_luis@lilly.com; Jordi Alonso - jalonso@imim.es; Rosa Lamarca - rlamarca@imim.es

* Corresponding author

Published: 28 July 2003

Health and Quality of Life Outcomes 2003, **1**:27

Received: 11 April 2003
Accepted: 28 July 2003

This article is available from: http://www.hqlo.com/content/1/1/27

Abstract

Background: Although health-related quality of life (HRQOL) instruments may offer satisfactory results, their length often limits the extent to which they are actually applied in clinical practice. Efforts to develop short questionnaires have largely focused on reducing existing instruments. The approaches most frequently employed for this purpose rely on statistical procedures that are considered exponents of Classical Test Theory (CTT). Despite the popularity of CTT, two major conceptual limitations have been pointed out: the lack of an explicit ordered continuum of items that represent a unidimensional construct, and the lack of additivity of rating scale data. In contrast to the CTT approach, the Rasch model provides an alternative scaling methodology that enables the examination of the hierarchical structure, unidimensionality and additivity of HRQOL measures. METHODS: In order to empirically compare CTT and Rasch Analysis (RA) results, this paper presents the parallel reduction of a 38-item questionnaire, the Nottingham Health Profile (NHP), through the analysis of the responses of a sample of 9,419 individuals.

Results: CTT resulted in 20 items (4 dimensions) whereas RA in 22 items (2 dimensions). Both instruments showed similar characteristics under CTT requirements: item-total correlation ranged 0.45–0.75 for NHP20 and 0.46–0.68 for NHP22, while reliability ranged 0.82–0.93 and 0.87–94 respectively.

Conclusions: Despite the differences in content, NHP20 and NHP22 convergent scores also showed high degrees of association (0.78–0.95). Although the unidimensional view of health of the NHP20 and NHP22 composite scores was also confirmed by RA, NHP20 dimensions failed to meet the goodness-of fit criteria established by the Rasch model, precluding the interval-level of measurement of its scores.

Introduction

Several questionnaires have been developed and are currently in extensive use to assess health-related quality of life (HRQOL) [1]. Such instruments may offer satisfactory properties in terms of measurement (i. e. validity and reliability), but their length often limits the extent to which they are actually applied in patient care. The availability of shorter instruments would prove highly advantageous in many situations, both in clinical practice and research: questionnaires may require excessive patient or interviewer time, or may be inappropriate if the patient is unable to participate in a lengthy procedure; in order to reduce the burden of response, shorter instruments might also prove beneficial when administered as part of a

multipurpose battery of different questionnaires, or when repeat assessments are required.

Efforts to develop short questionnaires have largely focused on reducing existing instruments. The methodology used to such ends has, to date, proved heterogeneous and lacking in standardization. The approach most frequently employed when seeking to shorten instruments seems to be statistical, and includes factor analysis, correlations between long and short-forms, correlations between item and composite scores, Cronbach's Alpha per scale, or stepwise regression [2]. These procedures all are based on the same underlying scaling model. The model, which could be called additive, assigns a measure, on a scale, as the sum of the responses to each item on the scale [3]. The additive model does not consider item hierarchy, and the criteria for the final selection are supplied by internal consistency checks. The additive model may be considered as the best exponent of Classical Test Theory (CTT) in test development and construction [3,4].

An alternative scaling approach, and reduction procedure, is a methodology based on the concept proposed by the Danish mathematician, Georg Rasch [5]. Built around a dichotomous logistic response model (suitable for *Yes/No* response choices) [6–8], Rasch specifies that each item response is taken as an outcome of the linear probabilistic interaction of a person's "ability" and a question's "difficulty" [5]. The Rasch model constructs a line of measurement with the items placed hierarchically and provides fit statistics to indicate just how well different items describe the group of subjects and how well individual subjects fit the group [9,10].

At all events, care must always be taken with respect to the possible weaknesses of the measurement properties of a shortened instrument [11]. Such weaknesses may be of particular importance with the additive model, since the number of items has an important influence on the final measurement properties of the questionnaire, especially with respect to reliability, and the form of score distribution (i. e., significant ceiling and floor effects) [12].

In order to empirically compare their results, the reduction of the Spanish version of the Nottingham Health Profile (NHP38) [13] was independently performed with CTT and Rasch Analysis. The measurement properties of the resulting questionnaires were tested and compared.

Monitoring the HRQOL of different populations demands global evaluations across a number of different health conditions and sociodemographic groups. In such a context the evaluator may require a single indicator or index number to describe the health status of the population being assessed. Thus, in both approaches, the items were selected in such a way so as to ensure that the reduced questionnaires would provide a unique summary index, indicating the health status of respondents to the questionnaire with a single number. Although a single number makes the results easier to use, not all developers or consumers of HRQOL measures accept the need for or desirability of summarizing health into a single index. A single health index cannot be a wholly comprehensive measure. Unless the analyst can ascertain the relative contribution of different domains to the overall index score, changes or trends in the index value are difficult to interpret [14]. As an alternative to the aggregated index, both reduction approaches also considered a profile structure (multiple numbers) to summarize the data collected by the new instruments.

Methods

The Nottingham Health Profile

The Nottingham Health Profile (NHP38) is a generic measure of subjective health status developed in Great Britain in the 1970s and extensively used in Europe [1]. It contains 38 items with a 'yes/no' response format, describing problems on six health dimensions (Energy, Pain, Emotional Reactions, Sleep, Social Isolation and Physical Mobility). The Spanish version of the questionnaire was obtained through a process of precise translation (using translation and back-translation procedures), aimed at achieving conceptual equivalence [13]. It has proved to be valid and reliable in several groups of patients [15]. The authors of the original version weighted each NHP38 item, to offset the differences in the scope of the problems described by each item. For each dimension (scale), the items were weighted by the paired comparison method proposed by Thurstone [16]. The NHP38 weighting has likewise been applied to the Swedish [17], French [18] and Spanish [19] versions of the questionnaire in order to assess cross-cultural equivalence and validate the process of adaptation. However, the use of an unweighted NHP38 scoring has been recommended for the Spanish version [19]. To such ends, the scores are obtained by adding together the number of affirmative answers for each scale in the questionnaire and expressing the number as a percentage, ranging from 0 (best health status) to 100 (worst health status).

Subjects

Data collection, intended for use in a common database covering all of the studies that have included the Spanish version of the NHP38 since its release in 1987, is described elsewhere [20,21]. The studies were identified by searches on Medline and the Spanish Medical Index from 1987 to 1995 (Key terms: Nottingham Health Profile, NHP, quality of life, measure of health status, questionnaire, reliability, validity, Spanish, and Spain). Other studies were identified from the Spanish NHP38 "cession

of use" registry, kept by one of the authors (JA) since 1987. Of the 119 studies identified, data were available from 45, covering a total of 9,419 individuals. The Spanish version of the NHP38 had been used in all the studies (all respondents reporting on their own HRQOL).

Selected variables from these 45 studies were collected in a common data base (i. e. responses to NHP38 items, gender, age, self-reported general health status, and study population).

Reduction based on Classical Test Theory (CTT)

The 38 items of the original Nottingham Health Profile (NHP38) were subject to item analysis, using standard statistical procedures [17,18]. The classical index of discrimination was obtained by calculating the corrected item-total correlation coefficients (r) for each item with its hypothetical scale [3]. Endorsement indices were also determined for each item by calculating the proportion (*p*) of people choosing to answer '*Yes*'. First of all, the NHP38 items with a *r* (<0.4) and a low (<0.20) or high *p* (>0.80) were excluded [22]. Exploratory Factor Analysis (EFA), employing Principal Axis Factor extraction and Promax rotation, was performed on the remaining items. EFA deleted all cases with missing values listwise (only cases with nonmissing values for all the items involved were used). A secondary reduction was then performed by deleting those items showing a low portion of the test score variance associated with the variance on the common factors (Communality < 0.3), as well as those items showing its highest factor loading on the main factor to be lower than 0.4, and those items with similar (difference ≤ 0.1) loadings on different factors.

Cronbach's alpha coefficient [23] was calculated on the scales (factors) resulting from the EFA, to estimate the internal-consistency reliability of each new composite score. Following the basics assumptions of CTT [3,4], a summary score of the reduced questionnaire was obtained by summing and averaging the scores of their component dimensions. The reliability of the summary score was estimated using the formula proposed by Nunnally and Bernstein (pp. 268) [3]. Additional EFA, based on principal component extraction, was used to determine whether the new dimensions could be reduced to a unique summary score.

Reduction based on Rasch analysis

Through log-odds, the Rasch model specifies that the probability of response of person n to item i is governed by location B_n for the subject (person measure) and location D_i for the item (item calibration), along a common continuum of measurement:

$$Log\,[P_{ni1}/P_{ni0}] = B_n - D_i$$

where, P_{ni1} is the probability of a "*Yes*" response to item i and P_{ni0} is the probability of a "*No*" response. When $B_n > D_i$, there is more than 50% chance of a "*Yes*" response. When $B_n = D_i$, the chance for a "*Yes*" response is 50%. When $B_n < D_i$, the probability is less than 50%. Each facet in the model (B, D) is a separate parameter. Estimates of one of the sets of parameters are not affected by the other. This mathematical property enables "test-free" and "person-free" measurement. This property implies that the parameter that characterize an item does not depend on the ability distribution of the examinees and the parameter that characterize a subject does not depend on the set of test items.

Item calibration defines the hierarchical order of severity ("difficulty") of the items along the health continuum. Item calibration is expressed in log-odd units (logits), positioned along a hierarchical scale. A logit is defined as the natural log of an odds ratio. Logits of greater magnitude represent increasing item severity. One logit is the distance along the health continuum that increases the odds of observing the event specified in the measurement model by a factor of 2.718, the value of e, the base of natural or Napierian logarithms used for the calculation of "log-" odds. All logits are the same length with respect to this change in the odds of observing the indicative event.

The unidimensionality of a scale can be evaluated by the pattern of item goodness-of-fit statistics and by a formal test of the assumption of local independence [5,9,10].

The original NHP38 was consecutively analyzed with the Rasch dichotomous response model. The Rasch analysis was performed with Version 2.7.3 of the BIGSTEPS computer program [25]. To avoid negative values, and to express the resulting scores on a 0 (best health status) to 100 (worst health status) scale score, the initial BIGSTEPS estimates were rescaled in all analysis, setting a new origin (49.73 units) and spacing (11.84 units/1 logit) for the scale [9]. In order to determine the precision of each estimate, an associated standard error (SE) was calculated for each item and person in the sample. The person separation index (PSEP) was also calculated. The PSEP is a ratio of standard deviation that describes the number of performance levels the test measures in a particular sample. It is equal to the square root of true person variance divided by the error variance due to person measurement imprecision (PSEP = (True Variance_N / Error Variance_N)$^{1/2}$. The test reliability (R) of the person separation index (PSEP) can be expressed as $R = (PSEP)^2/(1 + PSEP)^2$ [20,21]. Hence, the separation index has to exceed 2 (or 3) in order to attain the desired level of reliability of at least 0.80 (or 0.90). If statistically distinct levels of person ability are defined as ability strata with centers three measurement errors apart, then the PSEP can be translated into the

number of statistically distinct person strata identified by the test (Person Strata = [4 · PSEP + 1]/3). A Person Strata of, "3" (the minimum level to attain a reliability of 0.90) implies that three different levels of performance can be consistently identified by the test for samples like that tested.

Chi-square fit statistics were used to determine how well each NHP38 item contributed to defining a common health variable (Goodness-of-fit test) [9,10]. The most commonly used chi-squares are known as Oufit and Infit. They are reported as Mean-Squares (MNSQ), that is, the chi-square statistics divided by their degrees of freedom (so that they have a ratio-scale form with expectation 1 and range 0 to + ∝.). Outfit is based on the conventional sum of squared standardized residuals. If X is an observation, E its expected value based on Rasch parameter estimates, and σ^2 its modeled variance of expectation, then the squared standardized residual is: $z^2 = (X - E)^2/\sigma^2$. Oufit is $\Sigma (z^2)/N$, where N is the sum of the number of observations. Outfit is sensitive to unexpected responses made by persons for whom item i is far too "easy" or far too "difficult". Infit is an information-weighted sum in which each square residual is weighed by its variance (σ^2). Infit can be calculated as $\Sigma(z^2\sigma^2)/\Sigma (\sigma^2) = \Sigma (X - E)^2/\Sigma (\sigma^2)$. Since variance is smallest for persons furthest from items i, the contribution to Infit of their responses is reduced. An item with an Outfit or Infit MNSQ near 0 indicates that the sample is responding to it in an overly predictable way. Item Outfit or Infit MNSQ values of about 1 are ideal by Rasch model specifications, and indicate local independence. Items with Outfit or Infit MNSQ values greater than 1.3 are usually diagnosed as potential misfits to Rasch model conditions and considered for deletion from the assessed sequence (More information about this issue is provided by Smith et al. (1998)[24]). Successive Rasch analyses were performed until a final set of items satisfied the model fit requirements.

Since Rasch analysis places both persons and items along the same latent dimension, one can ask whether there is a substantial number of persons who actually do respond as predicted by the Rasch model. For this reason, person fit statistics, based on Infit and Outfit mean-square statistics, were also calculated for the new short-form obtained by the Rasch approach.

In order to minimize the loss of sensitivity of the new short questionnaire, two additional scoring options were taken into account. Considering previous experience with the questionnaire [15,26], the 38 items of the NHP38 were regrouped into two new, different scales before Rasch analysis was performed: a Physical scale (containing Energy, Pain and Physical Mobility dimensions) – 19 items – and a Psychological scale (containing Emotional Reactions, Sleep and Social Isolation) – 19 items. Separate Rasch analysis were performed with the Physical and Psychological scales. For this purpose, the item calibrations obtained when all items were analyzed together were used as anchor (fixed) values. The displacement (divergence) of the local estimate away from the anchored value was provided for each Physical and Psychological item (results not shown).

Comparisons of the two reduced versions

In order to perform a validation study of the stability of the results obtained by the two different strategies for the reduction of the questionnaire, the subjects in the initial common database were randomly divided into two independent sub-samples. The analysis described above was performed on sub-sample A (85%, n = 8,015), and independently repeated for sub-sample B (15%, n = 1,404)(15% was an arbitrary percentage which ensured that sub-sample B was representative of the age and study population sub-groups)

In order to compare the performance of the reduced versions, the following analyses were carried out: 1) Pearson and Spearman's coefficient of correlation was calculated comparing the original NHP38 and the CTT and Rasch analysis reduced scales; 2) Reliability estimates and item-total correlation coefficients were obtained for the Rasch analysis reduced scales and compared with the estimates obtained for the scales resulting form the CTT analysis; 3) the items and scales reduced by CTT were Rasch analyzed, and the results compared with those obtained by the Rasch reduction of the original questionnaire; 4) distribution patterns of scores and measures were described for each reduced questionnaire. Principal component extraction was also used to determine whether the Physical and Psychological Rasch scales could be reduced to a single summary score. The unidimensionality of the whole Rasch reduced version was further explored through the examination of the residual correlation matrix of a one-factor exploratory factor analysis of the items (Principal Axis Factor extraction).

Results

Table 1 shows the main characteristics of the population in the common database obtained from the 45 studies. The mean age of the overall sample was 57 (range 12 to 99). Nearly 50% of the sample were female. The subjects ranged from individuals from the general population to people suffering different clinical pathologies. Around 50% of the dataset comprised individuals from the general population. Among those suffering pathologies, diseases of musculoskeletal system and connective tissue were the most frequent.

Table 1: Characteristics of the study population

	ALL n = 9,419	MALES n = 4,478†	FEMALES N = 4,908†
Gender (%)		47.5	52.1
Age groups (%)			
12 – 44	24.5	23.1	26.0
45 – 54	14.9	15.5	14.5
55 – 64	18.9	21.6	16.5
65 – 74	24.6	25.6	23.7
75 – 99	16.7	14.1	19.2
Study populations			
General population	40.8	38.0	43.5
Primary care patients	7.4	4.5	10.0
Musculoskeletal system & connective tissue diseases	9.2	5.5	12.7
COPD/Asthma	9.2	14.0	4.8
Toxic Oil Syndrome	8.9	6.5	11.2
Chronic Kidney Failure	7.6	9.4	6.0
Cardiovascular diseases	3.2	5.5	0.9
Others	13.8	16.6	10.9

† For a subset of individuals (n = 33) information on gender was not available

Ten NHP38 items showing low r (<0.4) and low p (<0.20) values (range of p values was 0.09 to 0.56) were excluded in the first stage of the CTT approach (Table 2). The EFA of the 28 remaining items revealed a four-factor structure through the evaluation of the scree test. Data were missing on 826 people (out of the 85% sample) for this analysis, but the individuals removed did not differ systematically from the retained cases by age (mean difference = 2 years), gender or population group.

A second reduction, based on the EFA results, concluded in a new short-form containing 20 items (NHP20) and covering four different health dimensions (factors). Given the content of the items, the different dimensions were correspondingly named Physical, Emotional, Pain and Sleep. Like the original NHP38 score, scores for these scales were obtained by summing the number of affirmative responses to the items and expressing them as percentages, range 0–100 (best-worst health status). Standards of internal consistency reliability were well satisfied by all the dimensions (Alpha range: 0.82–0.84). Principal components results indicated that a single component was an optimal solution (loadings range 0.77–0.85), accounting for 67% of the total variance for the four scales of the NHP20 (results not shown). This outcome supports the calculation of a summary measure of the NHP20 as a simple addition of its four components. Cronbach's alpha for the NHP20 summary score was 0.94, only a hundredth lower than the alpha calculated for the NHP38 summary score.

The Rasch analysis of the 38 items of the NHP38 showed 9 misfitting items. Infit MNSQ statistics ranged from 0.78 to 1.30 (SD = 0.14) and outfit MNSQ ranged from 0.62 to 2.39 (SD = 0.41). Misfitting items in this, and subsequent analyses, were removed until no further improvement in fit requirements was found. Sixteen items were discarded in this process, reducing the initial questionnaire to 22 items (NHP22). There were 6,052 individuals (out of 8,015) susceptible to measurement in the Rasch analysis. A total of 2,412 individuals (out of 8,015) were not considered for the analysis since they reported a minimum (n = 1,361) or a maximum (n = 146) extreme score, or lacking responses for the whole questionnaire (n = 456). Missing responses were estimated (imputed) for those individuals who missed some of the items of the questionnaire -but not all of them- (n = 487 out of the 6,052 analyzed). Rasch model-based imputation was performed as part of the BIGSTEPS [25] calculation during the item calibration. The Rasch dichotomous model provides an expected value of response x_{ni} for each person (n) – item (i) encounter. The expected value (E_{ni}) falls between 0 and 1 and is given by $E_{ni} = \Sigma k \pi_{nik}$ where π_{nik} is person n's modeled probability of responding to item i in category k (0 or 1) [10]. The standard deviation of the Infit and Outfit MNSQ for the new reduced version fell to 0.09 and 0.24 respectively. The PSEP for the NHP22 was 2.08 (R = 0.81). The PSEP produces 3 statistically distinct person strata. In the calibration, items varied in severity from 25.15 to 76.11 units, with a standard error of 0.37 to 0.63. Eighteen of the 22 items fit to define a unidimensional variable

Table 2: Reduced NHP38 version obtained through Classical Test Theory (CTT): the NHP20

Original NHP38 items By dimension			*1st set of criteria for reduction: Discrimination (r) & Endorsement (p)*		*2nd set of criteria for reduction: Factor analysis**			**NHP20**
Dimension	No. items	α	*Items deleted as r < 0.40*	*Items deleted as P < 0.20*	*Items deleted as communality <0.30*	*Items deleted as main loading <0.40*	*Items deleted as difference between similar loadings ≤ 0.1*	No. Items remaining
Energy (EN)	3	.76	-	-	-	-	*EN2, EN3*	1
Pain (P)	8	.90	-	*P2*	-	-	*P4, P5*	5
Emotional Reactions (EM)	9	.82	*EM8*	-	*EM5, EM7*	-	-	6
Sleep (SL)	5	.81	*SL1*	-	-	-	-	4
Social Isolation (SO)	5	.78	-	*SO2, SO3 SO4, SO5*	*SO1*	-	-	0
Physical Mobility (PM)	8	.83	-	*PM1, PM3 PM8*	-	-	*PM6*	4
Summary Index	*38*	*.95*	*(28 items remaining, α = 0.94)*					*20 (α = .92)*

* *Principal Axis Extraction (4 factors) and Promax rotation (Factor intercorrelation range: 0.50 – 0.73)* NHP items are: EN1-I'm tired all the time; EN2-Everything is an effort; EN3-I soon run out of energy; P1-I have pain at night; P2-I have unbearable pain; P3-I find it painful to change position; P4-I'm in pain when I walk; P5-I'm in pain when I'm standing; P6-I'm in constant pain; P7-I'm in pain when going up/down stairs; P8-I'm in pain when I'm sitting; EM1-Things are getting me down; EM2-I've forgotten to enjoy myself; EM3-I'm feeling on edge; EM4-These days seem to drag; EM5-I lose my temper easily these days; EM6-I feel as if I'm losing control; EM7-Worry is keeping me awake at night; EM8-I feel that life is not worth living; EM9-I wake up feeling depressed; SL1-I take tablets to help me sleep; SL2-I'm waking in the early hours ...; SL3-I lie awake for most of the night; SL4-It takes me long time to get to sleep; SL5-I sleep badly at night; SO1-I feel lonely; SO2-I'm finding it hard to contact people; SO3-I feel there is nobody I am close to; SO4-I feel I am a burden to people; SO5 I'm finding hard to get on with people;PM1-I can only walk about indoors; PM2-I find it hard to bend; PM3-I'm unable to walk at all; PM4-I have trouble getting up/down stairs; PM5-I find it hard to reach for things; PM6-I find it hard to dress myself; PM7-I find it hard to stand for long; PM8-I need help to walk about outside.

according to Rasch specifications (Infit and outfit MNSQ < 1.3). The item calibrations of the NHP22, stratified by the Physical and Psychological sub-scales, are shown in Table 3 (see column labeled "Anchored measure"). Items are arranged from more to less severe health status within each scale. The standard error and fit statistics for these estimates are also shown in Table 3. Nine of the 11 Physical items and 10 of the 11 Psychological items fit to define unidimensional variables by themselves. The PSEP was 1.39 (R = 0.66), producing 2.2 statistically distinct person strata. For the Psychological scale, the PSEP was 1.24 (R = 0.61, Person strata = 2). The 3 misfitting items (PM1 and PM4 on the Physical and EM1 on the Psychological scale) were the same 3 out of 4 that misfitted in the calibration of all the 22 items described above. According to the Outfit statistics, there were a few unexpectedly high and low scores across individuals for these 4 items. Considering (1) that their extreme positions in the hierarchies are, nevertheless, conceptually valid and (2) that their exclusion substantially decreased the PSEP of the scales (even when combined in a single index), these misfitting items were finally retained.

Ninety-two percent of people in the sample was properly measured by the items of the NHP22 according to the Infit criterion (MNSQ < 1.3). When the same criterion was applied to the outfit MNSQ, the percentage of subjects properly measured was 80%.

Table 4 shows the final content of both reduced versions, the NHP20 obtained by the CTT approach and the NHP22 obtained by the Rasch analysis. The NHP22 short-form contains items from the six dimensions of the original NHP38. Social Isolation was the only dimension from the original questionnaire not represented in the NHP20. The new reduced versions share 13 common items, that is, 65% and 59% of the total content, respectively.

Both reduction strategies provided equivalent results when validation sub-sample B (n = 1,404) was analyzed instead of sub-sample A (results not shown, available upon request).

Table 5 shows the Spearman's correlation coefficient of the NHP38, NHP20 and NHP22 scales. When comparing the correlations (*r*) of the NHP20 and NHP22 and the original, higher coefficients were found when the comparisons included similar quality of life domains (i.e NHP38 Physical mobility with NHP20 Physical -*r* = 0.94-, or with NHP22 Physical -*r* = 0.93-). The correlations of total-

Table 3: Reduced NHP version obtained through Rasch Analysis: the NHP22

PHYSICAL SCALE				
ITEMS	*ANCHORED MEASURE*	*SE*	*INFIT MNSQ*	*OUTFIT MNSQ*
PM3-UNABLE TO WALK	76.1	.62	.82	.96
PM1-WALKING LIMITED	64.1	.51	1.03	1.35
P2-AWFUL PAIN	59.7	.47	.89	.92
PM6-HARD TO DRESS	55.7	.45	.83	.72
P8-PAIN SITTING	53.1	.44	.93	.90
PM5-HARD TO REACH	47.5	.42	.83	.74
EN3-OUT OF ENERGY	44.7	.41	.99	.99
P3-CHANGE PAIN	41.5	.41	.96	.97
P4-WALK PAIN	39.1	.41	.92	.91
PM2-HARD TO BEND	34.8	.41	.90	.93
PM4-STAIRS HARD	25.2	.43	.99	1.50
PSYCHOLOGICAL SCALE				
ITEMS	*ANCHORED MEASURE*	*SE*	*INFIT MNSQ*	*OUTFIT MNSQ*
SO5-PEOPLE HARD	68.6	.55	1.00	1.17
SO2-CONTACT HARD	60.6	.48	.96	1.03
SO4-I'M A BURDEN	58.6	.46	.98	1.02
EM4-DAYS DRAG	56.1	.44	.90	.89
EM6-NO CONTROL	55.0	.44	.90	.88
EM9-DEPRESSED	48.1	.41	.86	.81
EM2-JOY FORGOTTEN	44.9	.41	1.06	1.11
SL3-CAN'T SLEEP	44.0	.40	.93	.88
SL5-SLEEPS BADLY	40.8	.40	.89	.86
SL4-SLOW TO SLEEP	39.2	.40	1.02	1.08
EM1-GETTING ME DOWN	36.8	.43	1.20	1.34

NHP38 scores and total-NHP20 and total-NHP22 scores were identical and high (0.97). A high association was also observed between total NHP22 and total NHP20 scores (0.95), along with the expected pattern of correlations between their scales.

Principal component analysis (PCA) results (Table 6) confirmed the adequacy of averaging the scales of both reduced versions to obtain a single summary score for each. The PCA identified a main component (initial eigenvalues: 2.7, 0.6, 0.4, and 0.3) that accounted for 67.5% of the total variance of the CTT reduced version (NHP20). For the Rasch analysis reduced version (NHP22), the PCA also distinguished a main component (initial eigenvalues: 1.7 and 0.3) that accounted for 85% of total variance. The loadings of the scales for each instrument on its own main component were substantial: 0.77 to 0.85 for the NHP20 and 0.92 for the NHP22 scales. The NHP22 residual correlations found with a one-factor exploratory factor analysis showed very low magnitudes in absolute values (Median = 0.044; 75th Percentile = 0.079), suggesting that the one-factor model does fit the data, as well as the unidimensionality of the items of the NHP22.

Table 6 summarizes the distributional properties of the NHP20 and NHP22 scores, as well as the main CTT and Rasch analysis results. The NHP20 scales resulted in a higher number of missing scores than the NHP22 scales, but this is not surprisingly given that missing responses were imputed for the Rasch model (as part of the BIGSTEPS calculation) but not the CTT model. It should be noted that Rasch and CTT analyses were conducted on the same sample. Differences in the final number of individuals considered in each analysis were due to the idiosyncrasy of each calculation procedure. In any case, the number of "common" individuals in each analysis (n = 5,741) were, in my view, sufficient to provide stable and comparable results (e.g. the number of "common" individuals represents 94% of the Rasch analysis sample (n = 6,052), and 80% of the EFA analysis sample (n = 7,189). Neither the NHP20 nor the NHP22 showed a normal dis-

Table 4: Content of the reduced NHP versions

Original NHP38 dimensions	Classical Test Theory reduction NHP20				Rasch reduction NHP22	
	Emotional	Physical	Sleep	Pain	Physical	Psychological
Energy						
EN1 I'm tired all the time	X					
EN2 Everything is an effort						
EN3 I soon run out of energy					X	
Pain						
P1 I have pain at night				X		
P2 I have unbearable pain					X	
P3 I find it painful to change position				X	X	
P4 I'm in pain when I walk					X	
P5 I'm in pain when I'm standing						
P6 I'm in constant pain				X		
P7 I'm in pain when going up/down stairs		X				
P8 I'm in pain when I'm sitting				X	X	
Emotional Reactions						
EM1 Things are getting me down	X					X
EM2 I've forgotten how to enjoy myself	X					X
EM3 I'm feeling on edge	X					
EM4 These days seem to drag	X					X
EM5 I lose my temper easily these days						
EM6 I feel as if I'm losing control	X					X
EM7 Worry is keeping me awake at night						
EM8 I feel that life is not worth living						
EM9 I wake up feeling depressed	X					X
Sleep						
SL1 I take tablets to help me sleep						
SL2 I'm waking in the early hours ...			X			
SL3 I lie awake for most of the night			X			X
SL4 It takes me long time to get to sleep			X			X
SL5 I sleep badly at night			X			X
Social Isolation						
SO1 I feel lonely						
SO2 I'm finding it hard to contact people						X
SO3 I feel there is nobody I am close to						
SO4 I feel I am a burden to people						X
SO5 I'm finding hard to get on with people						X
Physical Mobility						
PM1 I can only walk about indoors					X	
PM2 I find it hard to bend		X			X	
PM3 I'm unable to walk at all					X	
PM4 I have trouble getting up/down stairs		X			X	
PM5 I find it hard to reach for things		X				
PM6 I find it hard to dress myself					X	
PM7 I find it hard to stand for long					X	
PM8 I need help to walk about outside						

(X) indicates the items included in each dimension of the reduced questionnaires *Items common to the NHP20 and NHP22 questionnaires are shown in italics*

Table 5: Association* of the original NHP38 and the two alternative short-forms: the NHP20 and NHP22

	NHP20					NHP22		
	Total score	Emotional	Physical	Sleep	Pain	**Total score**	Physical	Psychological
Total NHP38	.97	.83	.82	.73	.76	.97	.88	.88
Energy	.77	.72	.69	.48	.60	.77	.76	.65
Pain	.84	.59	**.80**	.51	**.91**	.82	**.88**	.61
Emotional Reactions	.79	**.92**	.52	.58	.53	.78	.59	**.85**
Sleep	.78	.56	.50	**.98**	.53	.73	.55	**.79**
Social Isolation	.54	.59	.39	.37	.39	.61	.47	.66
Physical Mobility	.82	.58	**.94**	.47	.65	.84	**.93**	.60
Total NHP20	-	-	-	-	-	-	-	-
Emotional	.83	-	-	-	-	-	-	-
Physical	.84	.56	-	-	-	-	-	-
Sleep	.76	.56	.48	-	-	-	-	-
Pain	.79	.56	.66	.51	-	-	-	-
Total NHP22	.95	.82	.82	.71	.76	-	-	-
Physical	.88	.63	**.91**	.52	**.80**	.91	-	-
Psychological	.87	**.87**	.58	**.78**	.59	.88	.65	-

* Spearman's Correlation Coefficients

Table 6: Distribution of scores and summary Classical Test Theory (CTT) and Rasch analysis results for the NHP20 and NHP22

	NHP20					NHP22		
	Total score	Emotional	Physical	Sleep	Pain	**Total score**	Physical	Psychological
Number of items	20	7	5	4	4	22	11	11
Principal components results								
Loadings of the first component*	-	0.70	0.68	0.60	0.72	-	0.92	0.92
Distribution of scores								
Valid observations	7,243	7,382	7,442	7,455	7,452	7,559	7,558	7,557
Mean	35.36	30.45	44.72	40.62	28.24	31.04	29.42	28.21
Standard deviation	29.33	31.46	38.10	39.34	36.50	23.84	28.06	27.40
25th Percentile	10	0	0	0	0	8.77	0	0
50th Percentile	30	14.29	40	25	0	28.52	23.56	24.22
75th Percentile	55	57.14	80	75	50	46.94	48.18	46.46
% 0 score	10.8	30.1	27.0	33.0	49.9	15.7	28.1	26.9
% 100 score	2.3	5.5	17.3	20.3	12.0	1.7	3.0	3.0
CTT analysis results								
Item-total correlation (range)	0.45–0.65	0.51–0.62	0.57–0.71	0.51–0.75	0.65–0.68	0.46–0.65	0.47–0.68	0.47–0.64
Reliability								
Cronbach's α	-	0.82	0.83	0.88	0.87	-	0.88	0.87
Linear combination	0.94	-	-	-	-	0.93	-	-
Rasch analysis results								
Person separation	2.17	0.74	0.32	0.00	0.00	2.08	1.39	1.24
Person reliability	0.82	0.35	0.09	0.00	0.00	0.81	0.66	0.61

* One component accounts of 67.5% of the total variance for the NHP20, and 85% of the total variance for the NHP22.

tribution of scores ($p < 0.001$) -results not shown-. Total NHP20 scores showed a lower floor effect than total NHP22 scores (10.8% vs. 15.7%). For the component dimensions of both reduced versions, ceiling effects were always lower than the maximum arbitrary value suggested (15%) for individual applications of health status instruments.

All of the correlation coefficients of each NHP22 item and its hypothesized scale exceeded a value of 0.4 (Table 6). Each of the NHP22 scales bordered on the minimum item internal-consistency reliability standard of 0.90 recommended when individual decisions are made with respect to specific test scores [3].

When Rasch analysis was applied to the NHP20, the results did not confirm the adequacy of the version, with respect to valid and reliable measurements. Although the NHP20 total scores seem to possess acceptable Rasch model properties, similar to those provided by the NHP22 total scores, its component scales (Emotional, Physical, Sleep and Pain) showed poor results (person strata range from 0 to 1.32, implying that, in the best of cases, only one level of performance could be consistently identified by the test), precluding its use under the Rasch model specifications.

Discussion

With a view to shortening the Nottingham Health Profile, two different approaches to item reduction were compared. The first approach was based on the successive statistical procedures of Classical Test Theory (CTT) [3,4], focusing on item difficulty (p) and discrimination (r) indices as well as exploratory factor analysis. The other approach was based on Rasch analysis [5,10]. The CTT approach produced a short version of 20 items (NHP20), describing problems on four health dimensions: Emotional, Physical, Sleep and Pain. The Rasch procedure generated a reduced version of 22 items (NHP22), measuring two different dimensions: Physical and Psychological. The content of the two was equivalent for 13 items (about 60% of total content).

While the NHP22 covered the entire range of dimensions considered by the original NHP38, the NHP20 eliminated (following the established "statistical" criteria) all the items in the Social Isolation sub-scale of the NHP38. Given that a component of the original scale has been eliminated, several questions may arise regarding the comparability of the new short-forms and the full version. Should the original factorial structure of the instrument be preserved when producing a short version of an established measure? Under what circumstances can modifications ignore the factorial structure of the original instrument? In this respect, Coste et al. [2] indicated that a preliminary issue to be addressed by the shortening process is to determine whether the original instrument should be considered as the reference. When the original instrument is considered as the "gold standard", the short-form should reproduce or predict the original instrument results. The high correlation (0.97) of the total scores of both short versions with the original instrument (NHP38), suggests that eliminating items did not cause a substantial change to the concept of perceived health status as measured by the NHP38. The pattern of correlation of the composite scales of the NHP20 and the NHP22 with the original dimensions of the NHP38, also indicates the convergence of results. In addition, the high association of the NHP20 and NHP22 scales (0.95 for summary and 0.78 to 0.91 for the related dimensions (NHP22 Physical and NHP20 Physical and Pain; and NHP22 Psychological and NHP20 Emotional and Sleep)) also suggests that both instruments are measuring comparable domains.

Seen from the perspective of the additive model of test construction, a preliminary conclusion, based on statistical findings, is that both reductions, NHP20 and NHP22 are good alternatives to the original NHP38. The assessed measurement properties of both questionnaires (including total and domain scales) are acceptable and similar to those described for the original version, suggesting that the two different methods used for the reduction, CTT and Rasch, have rendered two comparable versions of the original instrument that may be considered suitable for further testing in national studies.

To avoid criticism of the procedures chosen to examine the CTT approach, the decision was based on previously published studies [27]. Nevertheless, the somewhat arbitrary nature of the CTT analysis have to be explicitly acknowledged. The selection of items based on internal consistency indices may have led to items with excessive redundancy remaining, thereby reducing the breadth of measurement of the scale. Factor analysis is also controversial [28–30] since there is no single way to determine the number of factors to extract in the analysis. Problems related to component under- or over-extraction are frequent and lead to unreliable factor solutions, and therefore the inadequate choice of items [28–30]. It might also be argued that the use of standard factor analysis methods is inappropriate for dichotomous items. Phi correlation is a special case from the Pearson Product Moment correlation applied to data containing dichotomies [3] and is generated by the ordinary correlation formula generally used in factor analysis programs. As Gorsuch [29] indicated (p. 296), all the factor-analytic derivations apply to phi, "Factoring such coefficients is quite legitimate. Both phis and point biserials can be intermixed with product-moment correlations of continuous variables with no

major problems". On the other hand, other experts warn that factor analysis of dichotomous items can produce factors that reflect the distributions of the items more than the content of the items [3]. In any case, problems with factor-analyzing dichotomous items may be minimized by dropping items with low p-values prior to factoring.

Another important issue when discussing the additive model is the appropriateness of a summary index (or total scores) for the NHP22 and the NHP20. The problems with such indices are on two levels: practical and conceptual. From the practical point of view, and given the greater number of items involved in the calculation of total scores, summary indices were more affected by missing responses than the composite scales of each questionnaire. Nevertheless, the development of appropriate imputation techniques could solve the problem. On the other hand, one conceptual concern is that by using a single score, all the composite scales in the questionnaire are given the same weight. Thus some dimensions that should contribute less to the total variance are given the same importance as more powerful dimensions when calculating the total score. The latter (absence of unidimensionality) may not prove of concern here, given the results of the internal consistency and principal components analyses (PCA): the high internal consistency of the NHP20 and NHP22 total scores, estimated by item-total correlations, indicates that both shortening approaches succeeded in distinguishing homogeneous groups of items; moreover, when analyzing the NHP20 and NHP22 scales, the PCA identified only one principal component that accounted for most of the observed variance (the main component also had high and similar loadings with the subscales). Inspection of the residual correlations of a one-factor exploratory factor analysis also suggested the unidimensionality of the items of the NHP22 Rasch reduced version. In my opinion, the fact that there is only one component (dimension) for the NHP22 and NHP20 scores supports the use of the summary scores. Although this conclusion may appear to contradict a large body of previous research which suggests that physical and mental health form separate components [31] results from a recent study [32], aimed at testing the construct validity of the SF-36 in ten countries (including Spain), challenge the dichotomous conceptualization underlying scoring and interpretation. Structural equation modeling analyses supports the eight first-order factor model of health that underlies the scoring of the SF-36 scales, and two second-order factors that serve as the basis for summary physical and mental health measures. A single third-order factor was also observed "in support of the hypothesis that all responses to the SF-36 are generated by a single, underlying construct: health" [32]. The factor accounted for the correlation between physical and mental health factors, which in turn accounted for the correlations of the eight first-order factors and may, therefore be considered as the "cause" of all responses to the SF-36.

Through goodness-of-fit statistics and the investigation of the hierarchy of item calibrations, the unidimensional view of health of the NHP20 and NHP22 summary scales was confirmed by Rasch analysis. Thus, from the perspective of the Rasch model of test construction, the NHP20 and NHP22 total scales may also be considered as good substitutes for the original NHP38. Although the two scales (Physical and Psychological) of the NHP22 also met the Rasch model requirements [9,10], the four composite scales of the NHP20 failed to meet the minimum goodness-of-fit criteria of the analysis, thus undermining the validity of the measurements under a Rasch approach. The results, not surprisingly, given the way these dimensions were constructed, may indicate that better performance of the NHP20 could probably be achieved if it were made more comparable to the NHP22, for example, by developing two longer "physical" and "mental" scales, rather than 4 short ones (since reliability of person separation depends, to a certain extent, on the number of items in a scale).

The arbitrary decision that led to regroup the original NHP38 items into two different scales, before any Rasch analysis was performed, could be criticized, even more when the main goal of the study was intended to derive a scale that measured a single dimension of health. As indicated in the introduction, the "profile" structure was proposed in order to overcome the lack of comprehensiveness of a single index number. The decision was made on the basis of the experience with the questionnaire and the conviction that it adequately represents the physical-psychological duality inherent in any HRQOL measurement. On the other hand, using the item calibrations obtained when all items were analyzed together as "anchors" for these two scales is consistent with the intention that all NHP22 content should address a single underlying phenomenon (health).

Unfortunately, the results of the study do not provide strong evidence to prefer one of the instruments over the other. The only difference of note seems to be that the four component scales of the NHP20 did not fit the Rasch model specification. Since the goal was to develop a scale that reflects a single dimension of health and that can provide a single summary score, it seems that, from either a CTT or a Rasch model approach, the full NHP20 would be as acceptable as the full NHP22.

Even though the full NHP20 and NHP22 seem to share similar metric properties under both methodological approaches (CTT and Rasch), the poor Rasch performance of the four components of the NHP20 finally led to

choose the NHP22 as the short version of the original NHP38. Although the additive model, initially employed to score the dimensions of the NHP20, is very commonly used in HRQOL assessment, it has come under increasing scrutiny [33–35]. Two major conceptual limitations have been pointed out [33,35]: (1) the lack of an explicit ordered continuum of items that represent a unidimensional construct, and, (2), the lack of additivity of rating scale data, most often ordinal raw scores. The former implies that the greater the quantity, the larger the number associated with it. The latter condition indicates that the additional quantity associated with the increase of a number by one unit is of the same size, whatever the magnitude of the original quantity. In contrast to the additive approach, the Rasch model, provides a methodology that enables the examination of the hierarchical structure, unidimensionality and additivity of measures.

When evaluating an instrument with the Rasch model, more fundamental evidence may be provided to justify the use of scale scores on an interval level. Distances on the scales developed by the CTT approach are interpreted as equal over the full range of the scale. The scale is treated as an interval scale based on ordinal level item scoring [3]. This practice cannot be defended against the performance of measurement theories like the Rasch model. The Rasch scale is a statistically proven interval scale and is to be preferred. In addition to this argument, another advantage of using Rasch analysis with the NHP22 scales is that it deals with the missing data. As the Rasch algorithm compares each observed item score to an expected score, based on the overall scaling model, it uses expected score information when accounting for missing data. The procedure may offer a significant advantage when using the questionnaire at an individual level.

The analysis was based on a large heterogeneous group of individuals (including both healthy subjects from the general population and patients suffering diverse illnesses), leading to believe that findings may be generalized. Although these preliminary results suggest the adequacy of the new instrument, further research will be necessary to confirm the validity, reliability and stability of the item calibrations found for the NHP22. In any case, the initial conclusion is that the NHP22 questionnaire offers a promising short-form alternative to the original NHP38.

Author's contributions

LP conceived the study and managed its design, coordination and statistical analysis. JA and RL co-participated in all the process. All authors read and approved the final manuscript.

Acknowledgments

We are grateful to Ben Wright from the University of Chicago for his contribution to an early version of this paper.

The study has been supported by a grant from the "Fondo de Investigación Sanitaria (FIS), Expte. n° 96/0776". Additional support was received from the "Generalitat de Catalunya (CIRIT 1997 SGR 00359)".

References

1. McDowell I and Newell C: *Measuring Health: A guide to Rating Scales and Questionnaires* 2nd edition. *New York: Oxford University Press*; 1996.
2. Coste J, Guillemin F, Pouchot J and Fermanian J: **Methodological approaches to shortening composite measurement scales.** *J Clin Epidemiol* 1997, **50:**247-252.
3. Nunnally JC and Bernstein IH: *Psychometric theory* 3rd edition. *New York: McGraw-Hill*; 1994.
4. Crocker L and Algina J: *Introduction to classical and modern test theory Fort Worth: Harcourt Brace Jovanovich*; 1986.
5. Rasch G: *Probabilistic Models for Some Intelligence and attainment tests Chicago: Mesa Press*; 1993.
6. Blackwood L: **Latent variable models for the analysis of medical data with repeated measures of binary variables.** *Stat Med* 1988, **7:**975-981.
7. Douglas JA: **Item response models for longitudinal quality of life data in clinical trials.** *Stat Med* 1999, **18:**2917-2931.
8. Lindsey JK: **Directly modelling matched case-control data.** *Stat Med* 2000, **19:**35-44.
9. Wright BD and Stone MH: *Best Test Design: Rasch Measurement Chicago: MESA Press*; 1979.
10. Wright BD and Masters GN: *Rating Scale Analysis Chicago: MESA Press*; 1982.
11. Shrout PE and Yager TJ: **Reliability and validity of screening scales: effect of reducing scale length.** *J Clin Epidemiol* 1989, **42:**69-78.
12. Anastasi A and Urbina S: *Psychological Testing* 7th edition. *New Jersey: Prentice-Hall*; 1997.
13. Alonso J, Antó JM and Moreno C: **Spanish Version of the Nottingham Health Profile: Translation and Preliminary Validity.** *Am J Public Health* 1990, **80:**704-708.
14. Patrick DL and Erikson P: *Health status and health policy: Quality of life in health care evaluation and resource allocation New York: Oxford University Press*; 1993.
15. Alonso J, Prieto L and Antó JM: **The Spanish version of the Nottingham Health Profile: a review of adaptation and instrument characteristics.** *Qual Life Res* 1994, **3:**385-393.
16. McKenna SP, Hunt SM and McEwen J: **Weighting the seriousness of perceived health using Thurstone's method of paired comparisons.** *Int J Epidemiol* 1981, **10:**93-97.
17. Wiklund I, Romanus B and Hunt SM: **Self-assessed disability in patients with arthrosis of the hip joint. Reliability of the Swedish version of the Nottingham Health Profile.** *Int Disabil Stud* 1988, **10:**159-63.
18. Bucquet D, Condon S and Ritchie K: **The French Version of the Nottingham Health Profile. A Comparison of items Weights with Those of the Source Version.** *Soc Sci Med* 1990, **30:**829-835.
19. Prieto L, Alonso J, Viladrich MC and Antó JM: **Scaling the Spanish version of the Nottingham Health Profile: evidence of limited value of item weights.** *J Clin Epidemiol* 1996, **49:**31-38.
20. Prieto L, Lamarca R, Santed R, McFarlane D, Sanzo JM and Alonso J: **Reducing the items of the Nottingham Health Profile.** *Qual Life Res* 1997, **6:**703.
21. Prieto L, Alonso J, Lamarca R and Wright BD: **Rasch measurement for reducing the items of The Nottingham Health Profile.** *J Outcomes Meas* 1998, **2:**285-301.
22. Streiner DL and Norman GR: *Health measurement scales. A practical guide to their development and use Oxford: Oxford University Press*; 1989.
23. Cronbach LJ: **Coefficient alpha and the internal structure of tests.** *Psychometrika* 1951, **16:**297-334.
24. Smith RM, Schumacker RE and Bush MJ: **Using item mean squares to evaluate fit to the Rasch model.** *J Outcome Measurement* 1998, **2:**66-78.
25. Wright BD and Linacre JM: *User's Guide to BIGSTEPS: Rasch-Model Computer Program Chicago: MESA Press*; 1997.

26. Prieto L, Alonso J, Ferrer M and Antó JM: **Are results of the SF-36 Health Survey and the Nottingham Health Profile similar?: A comparison in COPD patients.** *J Clin Epidemiol* 1997, **50:**463-473.
27. Juniper EF, Guyatt GH, Streiner DL and King DR: **Clinical impact versus factor analysis for quality of life questionnaire construction.** *J Clin Epidemiol* 1997, **50:**233-238.
28. Wright BD: **Comparing Rasch measurement and factor analysis.** *Structural Equation Modeling* 1996, **3:**3-24.
29. Gorsuch RL: *Factor analysis* 2nd edition. *Hillsdale, New Jersey: LEA*; 1983.
30. Loehlin JC: *Latent variable models: an introduction to factor, path and structural analysis* 2nd edition. *Hillsdale, New Jersey: LEA*; 1992.
31. Hays R and Stewart A: **The structure of self-reported health in chronic disease patients.** *Psychol Assess* 1990, **2:**22-30.
32. Keller SD, Ware JE Jr, Bentler PM, Aaronson NK, Alonso J, Apolone G, Bjorner JB, Brazier J, Bullinger M, Kaasa S, Leplege A, Sullivan M and Gandek B: **Use of Structural Equation Modeling to Test the Construct Validity of the SF-36 Health Survey in Ten Countries: Results from the IQOLA Project.** *J Clin Epidemiol* 1998, **51:**1179-1188.
33. Coste J, Fermanian J and Venot A: **Methodological and statistical problems in the construction of composite measurement scales: a survey of six medical and epidemiological journals.** *Stat Med* 1995, **14:**331-345.
34. Haley SM, McHorney CA and Ware JE Jr: **Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale.** *J Clin Epidemiol* 1994, **47:**671-684.
35. McHorney CA, Haley SM and Ware JE: **Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. Comparison of relative precision using Likert and Rasch Scoring Methods.** *J Clin Epidemiol* 1997, **50:**451-461