Research Paper

PAIN

OPEN

Psychometric properties of Short Form-36 Health Survey, EuroQol 5-dimensions, and Hospital Anxiety and Depression Scale in patients with chronic pain

Riccardo LoMartire[a,b,*], Björn Olov Äng[a,b,c], Björn Gerdle[d], Linda Vixner[b]

Abstract

Recent research has highlighted a need for the psychometric evaluation of instruments targeting core domains of the pain experience in chronic pain populations. In this study, the measurement properties of Short Form-36 Health Survey (SF-36),EuroQol 5-dimensions (EQ-5D) and Hospital Anxiety and Depression Scale (HADS) were analyzed within the item response-theory framework based on data from 35,908 patients. To assess the structural validity of these instruments, the empirical representations of several conceptually substantiated latent structures were compared in a cross-validation procedure. The most structurally sound representations were selected from each questionnaire and their internal consistency reliability computed as a summary of their precision. Finally, questionnaire scores were correlated with each other to evaluate their convergent and discriminant validity. Our results supported that SF-36 is an acceptable measure of 2 independent constructs of physical and mental health. By contrast, although the approach to summarize the health-related quality of life construct of EQ-5D as a unidimensional score was valid, its low reliability rendered practical model implementation of doubtful utility. Finally, rather than being separated into 2 subscales of anxiety and depression, HADS was a valid and reliable measure of overall emotional distress. In support of convergent and discriminant validity, correlations between questionnaires showed that theoretically similar traits were highly associated, whereas unrelated traits were not. Our models can be applied to score SF-36 and HADS in chronic pain patients, but we recommend against using the EQ-5D model due to its low reliability. These results are useful for researchers and clinicians involved in chronic pain populations because questionnaires' properties determine their discriminating ability in patient status assessment.

Keywords: Chronic pain, Construct validity, EuroQol 5-Dimensions, Factor analysis, Hospital anxiety and depression scale, Internal consistency, Item response theory, Latent variable modeling, RAND-36, Short Form-36 Health Survey, Structural equation modeling, Structural validity

1. Introduction

Chronic pain is a globally prevalent condition that can permeate all aspects of a person's life.[55,61] Because it manifests itself differently in each individual, it is also notoriously difficult to measure. Considerable resources have therefore been invested into isolating core domains of the chronic pain experience.[16,24,35,64] Health-related quality of life (HRQoL) and emotional distress are 2 central domains that consistently recur in some form; they reflect perceived functioning and well-being in physical, mental, and social dimensions of health, and feelings of depression, anxiety, and anger, respectively. Both are recommended as core outcome domains in pain intervention clinical trials to increase research consistency.[16,35,64] Emotional distress has additionally been considered a critical psychosocial component in chronic pain patient profiling,[35] and was recently included as a central feature of the chronic pain experience in *ICD-11*.[41,62]

Unlike objective characteristics such as weight and height, human experiences are unobservable latent traits that need to be inferred from indicators through statistical procedures.[4] Self-administered questionnaires are the most commonly used indicators, with the Short Form-36 Health Survey (SF-36), the EuroQol 5-Dimensions (EQ-5D) for HRQoL,[7,33,66] and the Hospital Anxiety and Depression Scale (HADS) for emotional distress being the most predominant.[70] The measurement properties of these questionnaires have been extensively studied in various populations, including the general population, and patients with mental health conditions, cardiovascular disease, and cancer.[6,19,21] The combined evidence reveals that their properties vary widely in different populations,[6,19] and recent research has highlighted the paucity of psychometric evaluations of these instruments in chronic pain patients.[15,17]

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

[a] Division of Physiotherapy, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Huddinge, Sweden, [b] School of Education, Health and Social Studies, Dalarna University, Falun, Sweden, [c] Center for Clinical Research Dalarna - Uppsala University, Falun, Sweden, [d] Department of Medical and Health Sciences, Pain and Rehabilitation Centre, Linköping University, Linköping, Sweden

**Corresponding author. Address: Division of Physiotherapy, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Alfred Nobels allé 23, 141 83 Huddinge, Sweden. Tel.: +46 8 524 888 61; fax: +46 8 524 888 13. E-mail address: riccardo.lo.martire@ki.se (R. LoMartire).*

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

PAIN 161 (2020) 83–95

http://dx.doi.org/10.1097/j.pain.0000000000001700

This is an important void to fill because the accuracy with which a latent trait is captured depends on a questionnaire's validity.[57] For a questionnaire to be valid, it needs to be both based on solid underlying theory and have statistically robust empirical properties.[52,57] Rather than being an inherent questionnaire property, validity is a property of the test score, which, in turn, hinges on the population characteristics and the applied setting.[57] To avoid biased conclusions about the latent trait status, it is therefore critical to establish validity for individual populations separately.[57] Hence, psychometrically sound instruments are a prerequisite for the latent trait to be captured accurately, and the properties of SF-36, EQ-5D, and HADS have not yet been systematically addressed in chronic pain patients. We therefore evaluated their measurement properties in a large sample of this population.

2. Methods

2.1. Design and participants

This psychometric study was based on data from the Swedish Quality Registry of Pain Rehabilitation, which is an extensive nationwide database of patients with chronic musculoskeletal pain who are eligible for specialist rehabilitation.[43] It consists of self-report cross-sectional data from 35,908 patients who were consecutively recruited between 2009 and 2016 from 38 interdisciplinary pain specialist treatment clinics, distributed according to Swedish population density. These clinics target patients with particularly complex chronic pain conditions, characterized by psychological comorbidities, impaired work, and social ability, and a failure to respond to monodisciplinary interventions. Inclusion criteria were: aged 18 years or older with noncancer musculoskeletal pain for a minimum of 90 days. During their first visit, patients received information about the registry and signed a written informed consent form. They then provided information on demographics and pain characteristics, and completed the Swedish versions of HADS, EQ-5D, and SF-36 in this order. This study was approved by Uppsala's Medical Research Ethics Committee (DNR 2018/036).

2.2. Questionnaires

English translations of the questionnaires are provided in the supplementary materials (available at http://links.lww.com/PAIN/A877), along with path diagrams of their conceptual frameworks, which are also described below.

2.2.1. Short Form-36 Health Survey

SF-36 was designed to measure physical and mental health based on 8 health concepts: physical and social functioning, role limitations due to physical and emotional problems, mental health, vitality, bodily pain, and general health perception.[33,58,66,67] The scale was constructed to be suitable for use by anyone, irrespective of demographics or disease, and contains 36 items that are rated on 2 to 6 ordered categories.[67] SF-36 is often considered a measure of HRQoL,[15,17,33] due to the definitions of health and HRQoL being inconsistent but largely overlapping.[6,37] SF-36 is conceptualized as a hierarchical 2-level structure where the 2 constructs of physical and mental health (component summary scores), mediated through the 8 health concepts (subscales), drive the item responses.[33,65,67] Although this theory is largely consistent across studies, its empirical representations vary, which has resulted in various scoring methods and disparate association schemes linking the constructs, subscales, and items together.[33,63,66,68]

2.2.2. EuroQol 5-Dimensions

EQ-5D is a standard instrument in health economic evaluations that was constructed as a generic multiattribute utility measure to reflect the multidimensionality of HRQoL.[7,25] It contains 5 items with 3 ordered response categories each, which were selected to target different HRQoL dimensions. The original theory defined the items as independent causes of HRQoL and summarized them as a preference-based index.[7] However, because the items represent different aspects of HRQoL, they could simultaneously be conceptualized as indicators of a unidimensional construct.

2.2.3. Hospital Anxiety and Depression Scale

The Hospital Anxiety and Depression Scale is a 14-item questionnaire, rated on a 4-point ordinal scale, which was designed to measure emotional distress in nonpsychiatric patient populations.[70] The original theory defined 2 constructs of anxiety and depression represented by 7 items each;[70] however, alternative structures have since been proposed and evaluated in various populations.[2,19,42] These include a single distress construct,[47] variations of the 2 constructs of anxiety and depression,[40] 3 constructs of anxiety and depression combined with restlessness,[5,10] agitation,[28] or negative affectivity,[23] and bifactor structures with a general factor of emotional distress and 2 or 3 specific factors for residual item dependencies.[42]

2.3. Structural models

Structural models are empirical representations of a questionnaire's underlying theory that specify the number of latent traits and their relationship to each other as well as to the questionnaire items (**Fig. 1**).[48,52] The simplest structure is the unidimensional model, where all items measure a single latent trait. However, multidimensional models are required for questionnaires that measure more than one latent trait and can manage both between- and within-item multidimensionality. The correlated-traits model accommodates between-item multidimensionality and is implemented when subsets of items reflect different traits, which, in turn, are correlated with each other. By contrast, the two-tier model manages within-item multidimensionality in situations where items simultaneously measure some general traits of interest and specific features unique to item subsets, where the specific features can be thought of as residual dependencies not captured by the general trait.[12] The two-tier model simplifies to both the bifactor model, when a single general trait is measured,[30] and to the more widespread second-order model (not shown in **Fig. 1**), when items are constrained so that the general and specific loadings are proportional to each other.[53]

2.4. Functional models

Functional models describe how the latent trait and items relate to each other mathematically.[52] Item response theory (IRT) consists of a set of such generalized models that were developed for categorical data; they subclassify according to the mathematical function they are based on. Conceptually, IRT considers the item responses to be caused by (reflect) the latent trait. Item response theory–based analyses describe the shape of the item–trait relationship by estimating the probability of observed response patterns across the latent trait. Two common IRT models are the

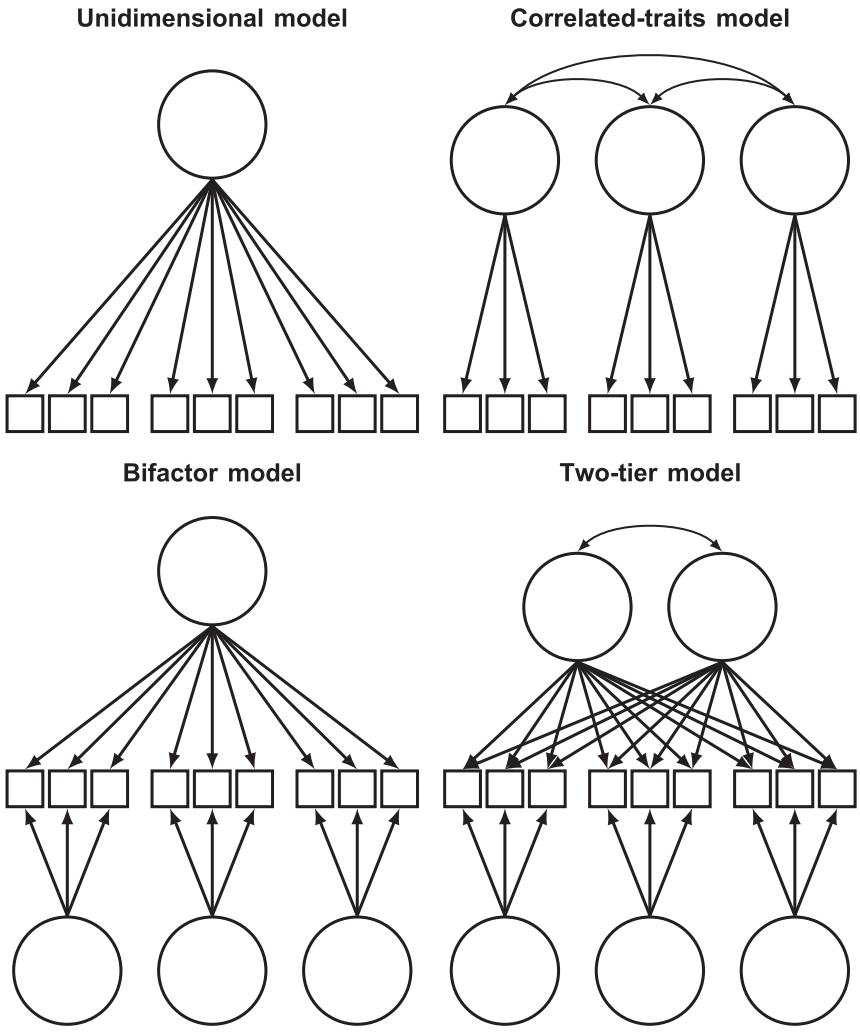

Figure 1. Path diagrams representing the structural models used in this study. Factors are represented by circles, items by squares, and causal pathways by arrows. Unidimensional model: all items load on one single factor that accounts for their covariance. Correlated-traits model: item subsets load on separate factors that are correlated; accounts for between-item multidimensionality. Bifactor model: each item loads on 2 uncorrelated factors; one general for all items and one specific for item subgroups; accommodates within-item multidimensionality. Two-tier model: a bifactor model with multiple general factors, which can be correlated.

logistic graded response model and its more constrained form, the 2-parameter logistic model; they are used for ordered and dichotomous item responses, respectively.[52,54] Item response theory models are advantageous in psychometric evaluations due to their flexibility and the detailed information they provide on item characteristics.[52,69] However, to be valid, they require that several assumptions are met: that the items are independent after accounting for the latent trait, that the item–trait relationship follows the specified mathematical function, and generally that the latent trait follows a predefined probability distribution.

2.5. Statistical analyses

Statistical analyses were computed in R (v3.5.2, R Core Team 2018) using the package "mirt" (v1.30).[14] Four types of structural models were used to accommodate the underlying theories of the questionnaires (**Fig. 1**), and IRT-based generalizations of the logistic graded response model (or the 2-parameter logistic model for the dichotomous items of SF-36) were used to describe the functional relationship between latent traits and item responses.[38,48,52,54] In confirmatory IRT models, the number of latent factors, which factors items can load on, and whether factors may intercorrelate are determined a priori, whereas the item parameters are estimated from the data. Models with normally distributed traits were estimated by full-information marginal maximum likelihood, using Bock and Aiken's expectation maximization algorithm and Cai's Metropolis-Hastings Robbins-Monro algorithm for up to 3, and 4 or more integral dimensions, respectively.[3,11] Model assumptions of the trait's normal distribution, item independence given the latent structure, and the item–trait relationship shape following the graded response model were assessed by: visual inspection of the test score and IRT score distribution, the item residuals, and comparing the observed and expected probabilities, respectively.[39,52] Path diagrams, assumption plots, the R scripts used in the analyses, and the models for scoring the questionnaires are provided in the supplementary materials (available at http://links.lww.com/PAIN/A877).

Patients with complete missing data were excluded from the analyses. For each questionnaire, k-fold cross-validation was computed to identify the model with the best fit and to assess the robustness of that model's parameter estimates.[32] Specifically, data were randomly split into 5 equal parts, models were fitted on 4 parts (training set), and model fit was evaluated using the

parameter estimates from the training set on the remaining part (validation set). Subsequently, models were refitted on the validation set using the same procedure with no additional constraints, and the training and validation set parameter estimate differences were computed. The procedure was repeated 5 times so that each part of the data was used as a validation set once.

Model selection was guided by several criteria. Firstly, after confirming model convergence, Akaike's and Schwartz's Bayesian information criteria were calculated to compare the model–model fit, with lower values supporting a better fit. Second, to evaluate the scale-level model-data fit, approximate fit indices based on the limited-information M_2* statistic were used, as it generally is unrealistic to expect exact model-data fit in IRT applications due to the many degrees of freedom resulting from the possible response pattern combinations.[13,39,52] The root mean square error of approximation (RMSEA) and the standardized root mean square residual (SRMSR) were the primary indicators, with estimates ≤ 0.05 considered an acceptable fit for both.[39,52] However, to facilitate comparison with previous studies, the Tucker–Lewis index (TLI) and the comparative fit index (CFI) were also included, which indicate the model fit relative to the null model, with estimates ≥0.90 typically considered acceptable. Third, item-level model-data fit was assessed based on the residual correlation matrix, with estimates ≤0.05 considered negligible,[39,52] and by comparing observed and expected S-X^2-based category response probabilities.[36] Fourth, person-level fit was assessed by the Zh-statistic, with estimates <|2.0| considered acceptable; negative and positive values suggest misfit (unpredictability) and overfit (redundancy), respectively.[50,52] Finally, to determine whether the model warranted its concomitant increase in complexity, expected and maximum a posteriori scores (IRT scores) for up to 2 and 3 or more latent dimensions, respectively, and their standard errors were compared across models.[52] For bifactor models, the explained common variance was also included as a measure of the general factor's strength relative to other factors, with estimates ≥0.85 suggesting that a unidimensional model would be sufficiently complex.[52]

Once the final model was selected, it was refitted on the complete data set to compute the final parameter estimates. The empirical internal consistency of each trait was also computed as a summary measure of precision,[14] with estimates ≥0.70 generally considered acceptable.[60] Finally, IRT scores were correlated with each other as well as the conventionally computed scores, as a measure of convergent and discriminant validity.

3. Results

3.1. Sample characteristics

Of the 35,908 patients, data from 34,910 (97.2%), 35,262 (98.2%) and 35,545 (99.0%) were used in the analyses, with partial missingness for 5703 (15.9%), 760 (2.1%), and 1946 (5.4%) of SF-36, EQ-5D, and HADS, respectively. Nearly 72% of the patients were female, the average age was 44 years, 84% had a secondary school education or higher, and 59% were actively employed. Moreover, 75% had a history of persistent or recurrent pain for 2 years or more, with the most prevalent locations after widespread pain being the lower back and the neck/shoulders, or upper extremities. **Table 1** details the sample characteristics.

3.2. Model selection

For all questionnaires, individual item distributions were either unimodal or monotone. The minimum number of observations per response category was 44 (followed by 171), 243, and 1239 for SF-36, EQ-5D, and HADS, respectively. No ceiling or floor effects were observed, with maximum and minimum test scores consistently being less than 0.1%.

Sixteen, 1, and 9 confirmatory models were computed for SF-36, EQ-5D, and HADS, respectively; all converged. For SF-36, scale-level indices supported that 5 models roughly had an acceptable fit. A two-tier model with all items free to load on 2 orthogonal general factors displayed the best fit, closely followed by its corresponding structure with 2 oblique general factors, whereas 3 bifactor models of the complete questionnaire and of the physical or mental health questionnaire sections separately had a somewhat worse fit. Piecewise indices more distinctly favored the orthogonal two-tier model because nonnegligible amounts of residual dependencies remained for the other models. However, the relationships between the general trait scores and conventional scores of all questionnaires revealed some trait–subscale inconsistencies and illogically high associations between the physical health trait and mentally oriented scales. The orthogonal two-tier model was therefore refitted with all factor loadings limited to positive values, which both mitigated the inconsistency and improved construct validity with no decrease in statistical properties. For EQ-5D, a unidimensional model was fitted, which showed satisfactory properties. Finally, indices unanimously supported that HADS was best modeled with a bifactor structure with 2 specific factors; other models performed similarly to each other, with the exception of the unidimensional model, which was markedly worse. For the selected models, latent trait distribution proxies supported normality, observed and expected categorical probabilities were similar, and item residual correlations were mostly within an acceptable level, thereby supporting the fact that the model assumptions were roughly met.

3.3. Short Form-36 Health Survey

Two dominant physical and mental health traits were observed, which accounted for 21.2% and 20.4% of the total variance, and constituted 30.3% and 29.3% of the variance common with the other latent traits, respectively. Scale-level indices unanimously supported an acceptable two-tier model fit {RMSEA: 0.041 (90% confidence interval [CI] 0.041-0.042); SRMSR: 0.038; TLI: 0.971; CFI: 0.976}. However, although mean item residual correlations were below 0.05, 72 individual correlations surpassed 0.05, with the maximum correlation of 0.19. Positive correlation patterns were observed both between items 24 and 25 from the mental health subscale with items 17 to 19 from the emotional role limitation subscale ($r \leq 0.13$), and between item 22 from the bodily pain subscale with items 13 to 16 from the physical role limitation subscale ($r \leq 0.13$). In addition, individual residual correlations remained, including between items 4 and 5 from the physical functioning subscale ($r = 0.15$), and items 23 and 27 from the vitality subscale ($r = 0.16$). Meanwhile, observed and expected response category probabilities generally matched well for most items, with the exception of items 29 and 31 of the vitality subscale, which showed distinct differences for higher categories. At person level, 4.0% (n = 1389) and 5.2% (n = 1808) of patients misfitted and overfitted the model, respectively. In addition, the cross-validation supported that most item parameters were relatively stable (loadings ≤|0.09|; intercepts ≤|0.6| logits); however, thresholds varied considerably more for items 10, 11, and 31 (≤|1.4| logits). Nonetheless, the variation in person IRT scores was less than the size of their standard errors for both traits (≤|0.3| logits).

Table 1

Sample characteristics.

Variable	Complete sample (n = 35,908)	SF-36 (n = 34,910)	EQ-5D (n = 35,262)	HADS (n = 35,545)
Females*	25,744 (71.7)	25,033 (71.7)	25,315 (71.8)	25,529 (71.8)
Age (y)†	44 (36-52)	44 (36-52)	45 (36-52)	44 (36-52)
Place of birth*				
Sweden	27,752 (77.3)	27,121 (77.7)	27,328 (77.5)	27,539 (77.5)
Europe, not Sweden	3059 (8.5)	2949 (8.4)	2980 (8.5)	3016 (8.5)
Outside Europe	4919 (13.7)	4668 (13.4)	4781 (13.6)	4814 (13.5)
Missing	178 (0.5)	172 (0.5)	173 (0.5)	176 (0.5)
Education level*				
Primary school	5852 (16.3)	5637 (16.1)	5710 (16.2)	5782 (16.3)
Secondary school	18,803 (52.4)	18,295 (52.4)	18,477 (52.4)	18,617 (52.4)
University	8692 (24.2)	8491 (24.3)	8564 (24.3)	8626 (24.3)
Other	2086 (5.8)	2034 (5.8)	2053 (5.8)	2055 (5.8)
Missing	475 (1.3)	453 (1.3)	458 (1.3)	465 (1.3)
Employment status*				
Employed	21,165 (58.9)	20,638 (59.1)	20,848 (59.1)	20,988 (59.0)
Student	1137 (3.2)	1099 (3.1)	1113 (3.2)	1124 (3.2)
Unemployed	10,607 (29.5)	10,268 (29.4)	10,365 (29.4)	10.480 (29.5)
Missing	2999 (8.4)	2905 (8.3)	2936 (8.3)	2.953 (8.3)
Pain characteristics				
Duration (y)†	5.4 (1.9-12.7)	5.4 (1.9-12.7)	5.4 (1.9-12.7)	5.4 (1.9-12.7)
NRS-10 past week pain intensity†	7 (6-8)	7 (6-8)	7 (6-8)	7 (6-8)
No. of pain locations (0-36)†	13 (7-20)	13 (7-20)	13 (7-20)	13 (7-20)
Primary pain location*				
Head	1803 (5.0)	1769 (5.1)	1780 (5.1)	1784 (5.0)
Neck, shoulders, and upper extremities	7739 (21.6)	7504 (21.5)	7590 (21.5)	7.667 (21.6)
Upper back and chest	1589 (4.4)	1554 (4.5)	1570 (4,5)	1575 (4.4)
Lower back	7014 (19.5)	6845 (19.6)	6910 (19.6)	6951 (19.6)
Abdomen	540 (1.5)	529 (1.5)	528 (1.5)	536 (1.5)
Hips and lower extremities	3465 (9.7)	3382 (9.7)	3417 (9.7)	3444 (9.7)
Widespread pain (varying)	11,909 (33.2)	11,564 (33.1)	11,686 (33.1)	11,804 (33.2)
Missing	1849 (5.1)	1763 (5.1)	1781 (5.1)	1784 (5.0)
Primary *ICD-10* diagnosis*‡				
Fibromyalgia (M79.7)	5532 (15.4)	5407 (15.5)	5433 (15.4)	5488 (15.4)
Unspecified pain (R52.9)	3382 (9.4)	3155 (9.0)	3348 (9.5)	3349 (9.4)
Myalgia (M79.1)	2843 (7.9)	2800 (8.0)	2810 (8.0)	2818 (7.9)
Low-back pain (M54.5)	2793 (7.8)	2730 (7.8)	2747 (7.8)	2767 (7.8)
Cervicobrachial syndrome (M53.1)	1986 (5.5)	1938 (5.6)	1948 (5.5)	1968 (5.5)
Lumbago with sciatica (M54.4)	1837 (5.1)	1788 (5.1)	1815 (5.1)	1820 (5.1)
Other chronic pain (R52.2)	1782 (5.0)	1745 (5.0)	1764 (5.0)	1768 (5.0)
Cervicalgia (M54.2)	1310 (3.6)	1274 (3.6)	1289 (3.7)	1290 (3.6)
Cervicocranial syndrome (M53.0)	1051 (2.9)	1031 (3.0)	1032 (2.9)	1042 (2.9)
Unspecified dorsalgia (M54.9)	649 (1.8)	626 (1.8)	635 (1.8)	644 (1.8)
Sequel of neck and trunk injuries (T91.8)	540 (1.5)	527 (1.5)	530 (1.5)	531 (1.5)
Pain in thoracic spine (M54.6)	443 (1.2)	438 (1.3)	438 (1.2)	438 (1.2)
Chronic intractable pain (R52.1)	389 (1.1)	378 (1.1)	376 (1.1)	386 (1.1)
Missing	1284 (3.6)	1247 (3.6)	1247 (3.5)	1262 (3.6)

* Frequency (percent).
† Median (25^{th} percentile and 75^{th} percentile).
‡ Primary *ICD-10* diagnoses shown for prevalence higher than 1%.
HADS, Hospital Anxiety and Depression Scale.

The two-tier model was also motivated with respect to parsimony because meaningful differences in the IRT-test score relationship and the IRT score standard errors were observed across models (**Fig. 2**). Both physical health and mental health were well targeted in patients with chronic pain, and had an internal consistency of 0.79 and 0.83, respectively. Items 10 to 11 and 24 to 25 best targeted patients with low health, whereas items 14 to 16, 18, 21, and 36 were best suited for patients with high health. Physical functioning, bodily pain, and role-physical items most strongly related to the physical health trait, whereas mental health, role-emotional, social functioning, and vitality items most strongly related to the mental health trait (**Table 2**). In addition, physical functioning items 3 to 5 and mental health items 26 and 30 were found to be particularly pure indicators of physical and mental health, respectively.

3.4. EuroQol 5-Dimensions

A single HRQoL trait was observed, which accounted for 35.5% of the total variance. All indices supported an acceptable unidimensional model fit, with scale-level indices within predefined thresholds (RMSEA: 0.046 [90% CI 0.040-0.052]; SRMSR:

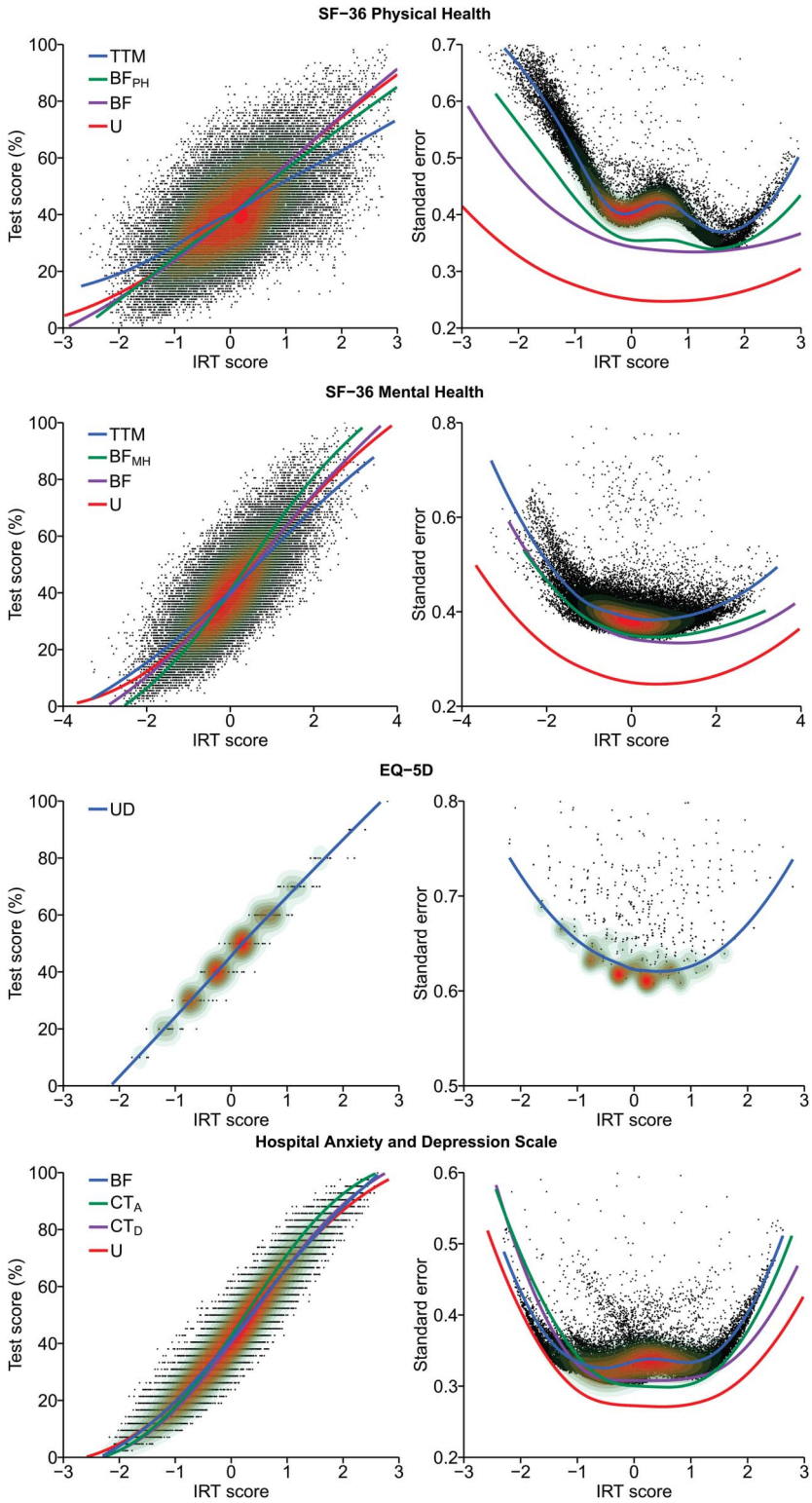

Figure 2. Relationship between IRT scores and standardized test scores (left), and IRT scores and their standard errors (right). The points with the overlaid contour plots show individual observations and sample density calculated from the final model, with higher density in red areas, whereas LOESS lines depict trends per model. Larger IRT scores indicate better health for SF-36, lower HRQoL for EQ-5D, and higher emotional distress for HADS. BF, bifactor model; CT, correlation traits model; HADS, Hospital Anxiety and Depression Scale; HRQoL, health-related quality of life; IRT, item response theory; TTM, two-tier model; U, unidimensional model.

0.031; TLI: 0.931; CFI: 0.972), item residual correlations below 0.06, observed and expected response category proportions matching well, and only 2.1% (n = 752) of patients misfitting the model. The cross-validation further sustained that item parameters (loadings ≤|0.06|; intercepts ≤|0.4| logits) and person IRT scores (≤|0.3| logits) were stable.

Table 2

Short Form-36 Health Survey parameter estimates.

Item	a_{G1} (SE)	a_{G2} (SE)	a_S* (SE)	d_1 (SE)	d_2 (SE)	d_3 (SE)	d_4 (SE)	d_5 (SE)	λ_{G1}	λ_{G2}	λ_S*	ECV_{G1}	ECV_{G2}	h^2
GH1 (1)	1.204 (0.052)	1.233 (0.051)	1.110 (0.019)[1]	1.010 (0.019)	−1.867 (0.022)	−4.020 (0.034)	−6.111 (0.058)		0.45	0.46	0.42[1]	0.34	0.36	0.59
PF1 (3)	2.569 (0.048)	0.000 (0.106)	0.000 (0.034)[2]	−3.424 (0.050)	−6.227 (0.083)				0.83	0.00	0.00[2]	1.00	0.00	0.69
PF2 (4)	3.102 (0.051)	0.167 (0.126)	0.308 (0.031)[2]	0.557 (0.024)	−5.172 (0.068)				0.87	0.05	0.09[2]	0.99	0.00	0.77
PF3 (5)	2.186 (0.030)	0.065 (0.089)	0.000 (0.024)[2]	0.446 (0.019)	−3.469 (0.036)				0.79	0.02	0.00[2]	1.00	0.00	0.62
PF4 (6)	1.972 (0.032)	0.366 (0.081)	1.591 (0.026)[2]	1.387 (0.023)	−2.070 (0.026)				0.64	0.12	0.52[2]	0.59	0.02	0.69
PF5 (7)	1.649 (0.031)	0.362 (0.069)	1.736 (0.027)[2]	4.037 (0.037)	0.206 (0.020)				0.56	0.12	0.59[2]	0.46	0.02	0.67
PF6 (8)	1.229 (0.019)	0.139 (0.051)	0.852 (0.018)[2]	0.998 (0.016)	−1.754 (0.019)				0.54	0.06	0.38[2]	0.67	0.01	0.44
PF7 (9)	2.244 (0.039)	0.469 (0.093)	2.161 (0.033)[2]	0.981 (0.025)	−2.397 (0.031)				0.63	0.13	0.60[2]	0.51	0.02	0.77
PF8 (10)	5.568 (0.273)	1.830 (0.250)	8.828 (0.424)[2]	12.728 (0.576)	2.164 (0.124)				0.52	0.17	0.82[2]	0.28	0.03	0.97
PF9 (11)	2.976 (0.095)	1.174 (0.130)	5.470 (0.129)[2]	9.939 (0.203)	3.481 (0.087)				0.45	0.18	0.83[2]	0.22	0.03	0.93
PF10 (12)	1.163 (0.023)	0.350 (0.049)	0.674 (0.017)[2]	3.591 (0.030)	0.346 (0.015)				0.53	0.16	0.31[2]	0.70	0.06	0.40
RP1 (13)	1.358 (0.051)	0.939 (0.062)	2.103 (0.059)[3]	−2.430 (0.047)					0.43	0.30	0.66[3]	0.26	0.12	0.71
RP2 (14)	1.814 (0.077)	1.404 (0.087)	2.525 (0.077)[3]	−5.253 (0.111)					0.48	0.37	0.66[3]	0.28	0.17	0.80
RP3 (15)	2.559 (0.079)	0.946 (0.111)	2.700 (0.084)[3]	−5.028 (0.114)					0.61	0.23	0.64[3]	0.44	0.06	0.84
RP4 (16)	2.285 (0.069)	1.004 (0.099)	2.340 (0.066)[3]	−4.566 (0.088)					0.60	0.26	0.61[3]	0.45	0.09	0.80
RE1 (17)	1.252 (0.137)	3.465 (0.104)	3.851 (0.104)[4]	−0.501 (0.040)					0.22	0.62	0.69[4]	0.06	0.42	0.91
RE2 (18)	1.800 (0.350)	7.116 (0.514)	8.076 (0.626)[4]	−4.329 (0.330)					0.16	0.64	0.73[4]	0.03	0.43	0.98
RE3 (19)	0.797 (0.091)	2.261 (0.050)	2.325 (0.043)[4]	−0.641 (0.027)					0.21	0.60	0.62[4]	0.06	0.46	0.79
SF1 (20)	1.376 (0.089)	2.179 (0.060)	1.863 (0.020)[5]	3.874 (0.035)	0.452 (0.024)	−2.218 (0.028)	−4.691 (0.040)		0.38	0.60	0.52[5]	0.19	0.47	0.78
BP1 (21)	1.522 (0.041)	0.888 (0.064)	1.769 (0.021)[6]	2.690 (0.028)	−1.620 (0.023)	−6.027 (0.054)	−7.824 (0.088)	−9.678 (0.169)	0.50	0.29	0.59[6]	0.37	0.13	0.68
BP2 (22)	2.122 (0.052)	1.116 (0.087)	1.769 (0.021)[6]	1.948 (0.027)	−1.990 (0.027)	−5.373 (0.048)	−8.154 (0.083)		0.62	0.33	0.52[6]	0.51	0.14	0.75
VT1 (23)	0.955 (0.065)	1.589 (0.044)	0.512 (0.018)[7]	0.454 (0.017)	−1.846 (0.021)	−3.559 (0.031)	−5.047 (0.044)	−7.172 (0.087)	0.37	0.62	0.20[7]	0.25	0.68	0.56
MH1 (24)	0.385 (0.069)	1.691 (0.026)	1.412 (0.026)[8]	5.133 (0.047)	3.186 (0.030)	1.786 (0.023)	0.481 (0.019)	−1.328 (0.020)	0.14	0.60	0.50[8]	0.03	0.57	0.63
MH2 (25)	0.746 (0.116)	2.887 (0.051)	2.280 (0.045)[8]	7.820 (0.101)	4.728 (0.063)	2.535 (0.040)	0.663 (0.028)	−1.944 (0.035)	0.18	0.70	0.55[8]	0.04	0.59	0.83
MH3 (26)	0.442 (0.096)	2.389 (0.031)	0.379 (0.024)[8]	2.406 (0.026)	−0.068 (0.020)	−1.698 (0.023)	−3.119 (0.030)	−5.855 (0.052)	0.15	0.80	0.13[8]	0.03	0.94	0.68
VT2 (27)	0.864 (0.076)	1.876 (0.042)	0.498 (0.018)[7]	0.086 (0.018)	−2.059 (0.023)	−3.586 (0.032)	−5.039 (0.044)	−6.855 (0.072)	0.32	0.69	0.18[7]	0.17	0.78	0.61

(*continued on next page*)

Table 2 (*continued*)

Item	a_{G1} (SE)	a_{G2} (SE)	a_S* (SE)	d_1 (SE)	d_2 (SE)	d_3 (SE)	d_4 (SE)	d_5 (SE)	λ_{G1}	λ_{G2}	λ_S*	ECV_{G1}	ECV_{G2}	h^2
MH4 (28)	0.560 (0.096)	2.385 (0.034)	1.636 (0.028)[8]	5.448 (0.050)	2.829 (0.031)	0.993 (0.023)	−0.690 (0.023)	−3.663 (0.036)	0.16	0.70	0.48[8]	0.04	0.66	0.75
VT3 (29)	0.682 (0.050)	1.190 (0.033)	1.532 (0.028)[7]	1.920 (0.025)	−0.082 (0.018)	−1.433 (0.022)	−2.469 (0.028)	−3.831 (0.038)	0.26	0.45	0.57[7]	0.11	0.34	0.59
MH5 (30)	0.516 (0.094)	2.336 (0.032)	0.273 (0.024)8	3.050 (0.029)	0.164 (0.019)	−1.674 (0.023)	−3.368 (0.031)	−6.161 (0.057)	0.17	0.79	0.09[8]	0.05	0.94	0.67
VT4 (31)	0.844 (0.069)	1.551 (0.057)	2.827 (0.105)[7]	1.042 (0.037)	−1.703 (0.052)	−3.765 (0.103)	−5.252 (0.140)	−6.860 (0.178)	0.23	0.41	0.76[7]	0.06	0.22	0.79
SF2 (32)	1.456 (0.085)	2.062 (0.062)	1.863 (0.020)[5]	4.712 (0.040)	1.132 (0.025)	−1.977 (0.027)	−4.634 (0.039)		0.41	0.58	0.52[5]	0.22	0.43	0.77
GH2 (33)	0.172 (0.025)	0.546 (0.015)	0.729 (0.016)[1]	2.548 (0.021)	1.368 (0.015)	0.313 (0.013)	−0.641 (0.013)		0.09	0.28	0.38[1]	0.03	0.35	0.23
GH3 (34)	0.825 (0.043)	0.996 (0.038)	1.718 (0.027)[1]	1.246 (0.021)	−0.711 (0.019)	−2.147 (0.026)	−3.696 (0.037)		0.30	0.36	0.63[1]	0.15	0.21	0.61
GH4 (35)	0.402 (0.031)	0.712 (0.021)	0.693 (0.015)[1]	2.562 (0.021)	1.188 (0.015)	−1.196 (0.015)	−2.182 (0.019)		0.20	0.35	0.34[1]	0.14	0.44	0.28
GH5 (36)	1.375 (0.078)	1.803 (0.065)	2.587 (0.054)[1]	−0.127 (0.026)	−2.869 (0.049)	−4.788 (0.074)	−7.990 (0.121)		0.36	0.47	0.67[1]	0.16	0.27	0.80

a, slope/discrimination; d, category intercept/threshold; λ, loading/correlation; h^2, variance explained/communality; G, general factor; S, specific factor; *, number of specific factor.
ECV, explained common variance.

EQ-5D was well targeted and provided most information in the central range of the latent trait (**Fig. 2**); however, with an internal consistency of 0.60, its precision was rather low. Item 4 best targeted patients with high HRQoL, whereas items 1 and 2 best matched those with low HRQoL. Finally, item 2 had the strongest relationship with the HRQoL trait, whereas item 5 had the weakest (**Table 3**).

3.5. Hospital Anxiety and Depression Scale

A strong emotional distress trait was identified, which represented 46.4% of the total variance, and constituted 74.1% of the variance common with the other latent traits. All indices supported an acceptable bifactor model fit, with scale-level indices within predefined thresholds (RMSEA: 0.048 [90% CI 0.046-0.049]; SRMSR: 0.032; TLI: 0.969; CFI: 0.983), and piecewise indices limited to minor misfits. At item level, mean residual correlations were below 0.05, and although 12 individual correlations surpassed that threshold, the maximum correlation was 0.09. Observed and expected response category probabilities also matched well, with deviances for items 2 and 9 in particular. At person level, 3.1% (n = 1105) misfitted the model and 1.9% (n = 678) overfitted it. The cross-validation further supported that item parameters (loadings ≤|0.06|; intercepts ≤|0.4| logits) and person IRT scores (≤|0.2| logits) were stable.

The bifactor model was also motivated with respect to parsimony because less complex models, although showing similar IRT-test score relationships, tended to overestimate IRT score precision (**Fig. 2**). In addition, the results showed that HADS was well targeted to the emotional distress levels in patients with chronic pain, with an internal consistency of 0.88. Items 1 and 8 best targeted patients with low emotional distress, whereas items 4 and 10 best targeted those with high emotional distress. All items were at least moderately related to the emotional distress trait, with anxiety items typically having the strongest relationship (**Table 4**). Finally, items 1, 7, 11, and 14 distinguished themselves as particularly pure indicators of emotional distress.

3.6. Latent trait relationships

The IRT score relationships were consistent with their underlying theories (**Fig. 3**). Physical health scores generally had the strongest associations with each other (SF-36 physical health vs SF-36 physical functioning subscale and SF-36 bodily pain subscale: r = |0.63-0.85|) and the weakest association with mental health scores (SF-36 physical health vs HADS and SF-36 mental health subscale: r = |0.20|). Likewise, mental health scores generally had the strongest association with each other (SF-36 mental health vs HADS and SF-36 mental health subscale: r = |0.72-0.89|; HADS vs SF-36 mental health subscale: r = −0.77), and the weakest association with physical health scores (SF mental health vs SF-36 physical functioning subscale: r = 0.13; HADS vs SF-36 physical functioning subscale: r = −0.24). The pattern was less obvious for EQ-5D, which was more equally associated to both physical (SF-36 physical health and SF-36 physical functioning subscale: r = |0.61|) and mental measures (SF-36 mental health, HADS, and SF-36 mental health subscale: r = |0.37-0.42|). Finally, all IRT scores were strongly related to their conventionally calculated scores ($r \geq$ |0.80|), most noticeably for the EQ-5D index and the HADS anxiety subscale, which had nearly perfect associations ($r \geq$ |0.92|).

4. Discussion

SF-36, EQ-5D, and HADS are widely used questionnaires for measuring HRQoL and emotional distress in the health sciences, and many theories pertaining to their latent structures have arisen over time. We compared the empirical representations of the most recognized theories, and evaluated their properties through a robust statistical procedure, in a large sample of chronic pain patients. For all 3 questionnaires, we identified conceptualizations that were structurally sound and logically associated, which provides evidence for their validity in the chronic pain population. However, because the internal consistency of our EQ-5D representation was unacceptably low, we recommend that our models be used to score SF-36 and HADS only.

Table 3

EuroQol 5-Dimensions parameter estimates.

Item	a (SE)	d1 (SE)	d2 (SE)	λ	h^2
Mobility (1)	1.292 (0.027)	0.606 (0.015)	−5.587 (0.068)	0.60	0.37
Self-care (2)	1.671 (0.039)	−1.889 (0.029)	−5.936 (0.079)	0.70	0.49
Activities (3)	1.301 (0.024)	1.943 (0.022)	−1.530 (0.019)	0.61	0.37
Pain (4)	1.306 (0.027)	5.795 (0.073)	0.712 (0.016)	0.61	0.37
Anxiety (5)	0.803 (0.017)	1.697 (0.017)	−1.564 (0.016)	0.43	0.18

a, slope/discrimination; d, category intercept/threshold; λ, loading/correlation; h^2, variance explained/communality.

Consistent with the original theory behind SF-36, we observed 2 general constructs of physical and mental health.[33,67] They were most accurately captured in a model that also accounted for unrelated item dependencies within the 8 SF-36 subscales. Although these constructs are generally agreed upon,[1,26,31,33,63,65,67,68] 2 conflicting perspectives, either viewing them as independent or correlated, have largely dominated the psychometric literature.[26,31,33,67] Our results support the former view owing to a better data fit and higher parsimony. Summary scores based on the independent representation have received criticism for negative scoring weights that result in inconsistencies between summary scales and subscales.[59] To a lesser extent, these concerns apply to our results because the physical and mental health scores were inversely associated with some SF-36 items. These inconsistencies were not eliminated by allowing the physical and mental traits to correlate, in line with previous findings.[26,31] Instead, the problem was mitigated to a higher degree by restricting item loadings to positive values on all latent factors. In practice, the inconsistency is primarily a concern towards the trait score extremes and the subscale scores should therefore be considered when interpreting the overall scores in this range.[59] However, although the selected model had the comparatively best properties of the fitted models, further refinements are likely possible. In agreement with previous reports,[1,18,26,31,65] the physical and mental health constructs were most strongly associated with the physical functioning and mental health subscales, respectively, whereas the social functioning, vitality, and general health subscales were more evenly associated to both constructs. Largely, they were also logically correlated with other scales, particularly so between the mental health trait and HADS, which both measure mental state. However, given that the physical and mental health traits were uncorrelated by design, a small albeit seemingly contradictory association arose between the physical health construct and mentally oriented scales. A potential explanation for this is the importance of physical health perception for mental health, as previously motivated for the association between the physical health summary score and the mental health subscale of SF-36.[49] Hence, in general agreement with most previous studies, our results support 2 summary scales and 8 subscales, but suggest that the summary scales are independent rather than correlated, which is still an open question in the scientific community.[1,26,31,33,63,65,67]

Despite the conventional conceptualization of EQ-5D as a multidimensional scale, its HRQoL construct was acceptably represented by a unidimensional model. It is justifiable to model the scale both ways from the perspective of dimensionality alone because the unidimensional model only assumes that items share one common construct, not whether systematic components unique to each item remain within the residuals.[51] However, from a causality perspective, it may be conceptually inappropriate to model EQ-5D within the IRT framework because the original theory defines the items as independent causes of HRQoL, whereas they are viewed as correlated effects of HRQoL in IRT. Nonetheless, EQ-5D's causal structure is an open question, with recent studies providing some support for that it contains both cause and effect items.[27,29] Our results are coherent with other IRT evaluations in chronic pain and mental illness patients, which also assessed the properties of EQ-5D as a unidimensional scale and reported acceptable unidimensionality.[34,46,56] Our IRT-based score was strongly associated with the preference-based EQ-5D index and other physical scales and, to a lesser extent, with mental scales. These findings support construct validity and are intuitive because 4 of 5 items in EQ-5D target physical

Table 4

Hospital Anxiety and Depression Scale parameter estimates.

Item	a_g (SE)	a_s* (SE)	d_1 (SE)	d_2 (SE)	d_3 (SE)	λ_G	λ_S*	ECV_G	h^2
Tense (1)	1.789 (0.019)	0.286 (0.021)[1]	2.650 (0.023)	0.231 (0.016)	−1.639 (0.019)	0.72	0.12[1]	0.98	0.53
Enjoy (2)	1.294 (0.018)	1.431 (0.020)[2]	1.925 (0.021)	−1.284 (0.019)	−3.041 (0.027)	0.50	0.56[2]	0.45	0.56
Frightened (3)	2.501 (0.031)	1.547 (0.035)[1]	1.206 (0.024)	−2.076 (0.029)	−4.500 (0.046)	0.74	0.46[1]	0.72	0.75
Laugh (4)	2.105 (0.025)	1.823 (0.026)[2]	1.880 (0.026)	−1.874 (0.026)	−5.994 (0.057)	0.65	0.56[2]	0.57	0.73
Worried (5)	2.603 (0.029)	1.240 (0.028)[1]	2.866 (0.031)	−0.191 (0.022)	−2.628 (0.029)	0.78	0.37[1]	0.82	0.74
Cheerful (6)	2.076 (0.023)	1.481 (0.022)[2]	1.737 (0.023)	−1.405 (0.022)	−4.431 (0.038)	0.68	0.48[2]	0.66	0.69
Relaxed (7)	2.782 (0.060)	−0.981 (0.054)[1]	4.478 (0.084)	1.252 (0.032)	−3.196 (0.062)	0.82	−0.29[1]	0.89	0.75
Slowed down (8)	1.343 (0.016)	0.677 (0.015)[2]	3.583 (0.028)	0.729 (0.015)	−1.281 (0.016)	0.59	0.30[2]	0.80	0.44
Frightened (9)	2.310 (0.027)	1.321 (0.028)[1]	1.257 (0.023)	−2.034 (0.026)	−4.829 (0.045)	0.73	0.42[1]	0.75	0.71
Lost interest (10)	1.423 (0.017)	0.832 (0.016)[2]	0.877 (0.016)	−1.347 (0.017)	−3.821 (0.030)	0.60	0.35[2]	0.74	0.48
Restless (11)	1.275 (0.015)	0.128 (0.018)[1]	1.736 (0.017)	−0.091 (0.014)	−2.207 (0.019)	0.60	0.06[1]	0.99	0.36
Anticipation (12)	2.474 (0.033)	2.497 (0.039)[2]	3.727 (0.046)	−0.069 (0.025)	−4.695 (0.055)	0.63	0.64[2]	0.50	0.81
Panic (13)	2.360 (0.026)	1.036 (0.025)[1]	0.906 (0.021)	−1.713 (0.024)	−4.309 (0.038)	0.76	0.34[1]	0.84	0.70
Enjoy Book (14)	1.638 (0.019)	0.606 (0.016)[2]	1.026 (0.017)	−1.338 (0.018)	−2.801 (0.024)	0.67	0.25[2]	0.88	0.51

a, slope/discrimination; d, category intercept/threshold; λ, loading/correlation; h^2, variance explained/communality; G, general factor; S, specific factor; *, number of specific factor.
ECV, explained common variance.

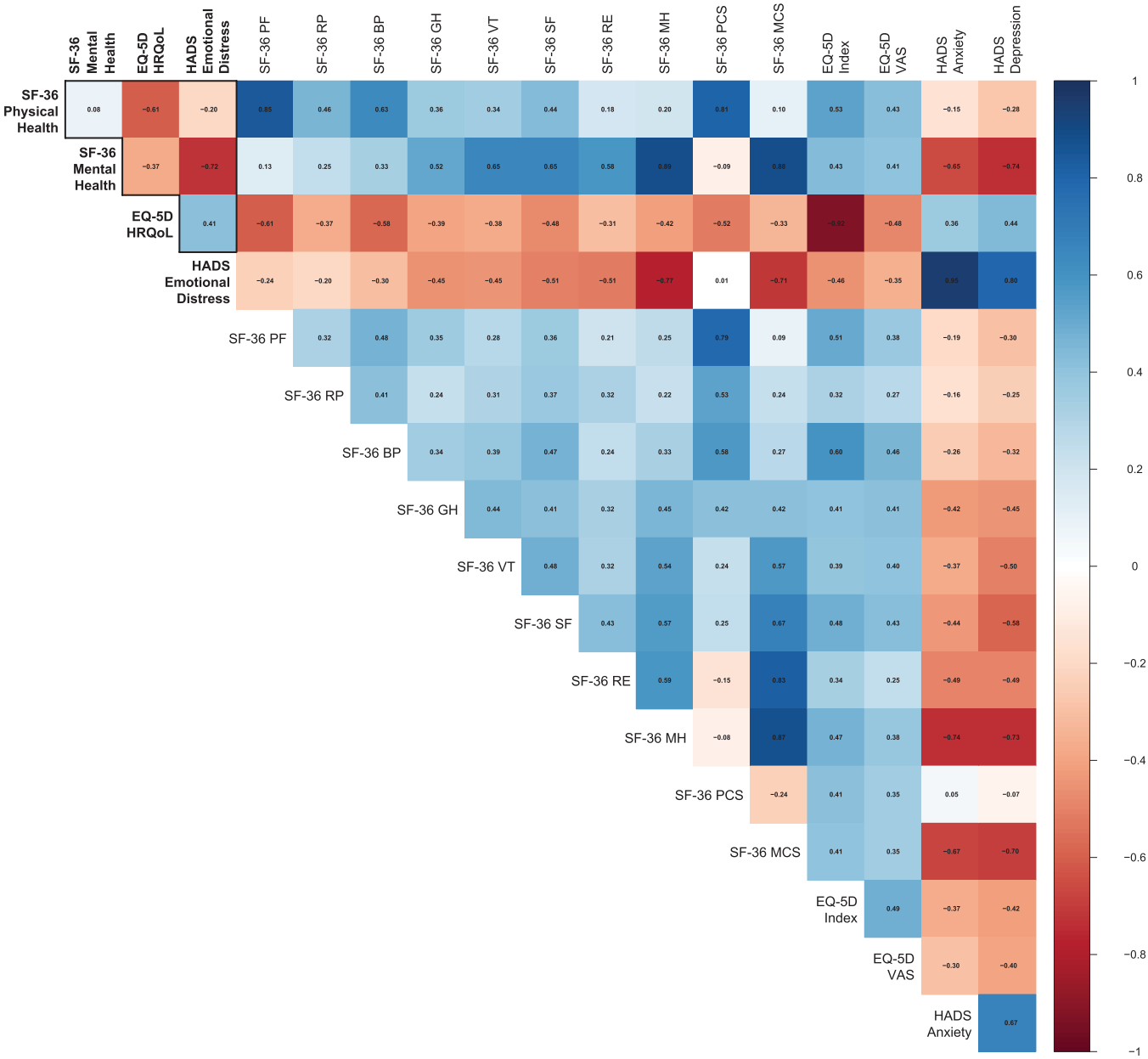

Figure 3. Associations between the IRT scores from the final models (bold) and the conventional scores for SF-36, EQ-5D, and HADS. Larger IRT scores indicate better health for SF-36, lower HRQoL for EQ-5D, and higher emotional distress for HADS. Conventional scores were calculated based on Ware et al., 1993, the United Kingdom time trade-off value set, and as item summaries for SF-36, EQ-5D, and HADS, respectively. BP, bodily pain; GH, general health; HADS,Hospital Anxiety and Depression Scale; HRQoL, health-related quality of life; IRT, item response theory; MH, mental health; PCS and MCS, physical and mental component summary scores, respectively; PF, physical functioning; RE, role-emotional; RP, role-physical; SF, social functioning; VT, vitality. n = 31,050.

health.[7] They are also consistent with studies of other populations, which have reported similar associations.[6,9] Regardless, we do not recommend the use of the IRT-based score because its internal consistency was too low to reliably discriminate between patients of different HRQoL levels. This is consistent with previous results,[34,56] and is hardly surprising because items intentionally target different HRQoL aspects. Interestingly, the multidimensional 3-level version of EQ-5D is also known to have low discriminatory power, resulting in the development of a 5-level version.[8] Our results are limited to the former, and it is possible that an IRT score of the latter would have acceptable reliability.

Although HADS has received considerable criticism in the past,[20] our results supported that it is a valid and reliable questionnaire. However, rather than the 2 original constructs of anxiety and depression, we observed one strong emotional distress construct. Similar to SF-36, the construct was most accurately captured in a model that accounted for unrelated item dependencies within the anxiety and depression subscales. This result corresponds to those of another study that compared the same latent structures in a meta-confirmatory factor analysis of 28 samples from various populations.[42] Conversely, 2 studies that evaluated the dimensionality of HADS in one sample of musculoskeletal pain patients came to somewhat different conclusions, which nonetheless are supported by our findings.[44,45] The first found that 2 highly correlated (r = 0.80) constructs of anxiety and depression had an acceptable fit with item 7 excluded.[44] Similarly, we observed a nearly acceptable scale-level fit for the original 2-factor model in a sample subgroup,

and item 7 also behaved differently in our analysis by loading negatively on the specific factor. Their second study suggested that HADS was sufficiently unidimensional to be used as a global measure of emotional distress, but that residual patterns related to the anxiety and depression subscales of a considerable size remained.[45] Our findings are largely in agreement because we found additional features not accounted for by the emotional distress construct that was necessary to factor into the model. The combined evidence thereby supports our selected structure for HADS.

For all questionnaires, the latent structures selected had the best statistical properties. However, inspecting the fit of individual SF-36 items revealed that small residual dependencies, not accounted for by the model, remained. Such dependencies are common and may reflect interpretation difficulties or response bias,[1,49] and are unlikely to implicate consequences other than a slight overestimation of the latent trait reliability. In addition to model fit, cross-validation supported that the item parameters obtained were generally stable for all questionnaires, although a more pronounced instability was observed for 3 SF-36 items, which we attribute to data sparsity owing to the many possible response patterns. Nevertheless, the IRT scores remained stable, which supports the fact that our models capture the marginal relationships in this population and that the models provided can be used to score HRQoL and emotional distress in chronic pain patients. EQ-5D was unidimensionally modeled and its scoring is thus unequivocal, but to accurately score the multidimensionally modelled traits, it is necessary to estimate them jointly with all modeled factors. Trait scores can be derived in accordance with recommended methods,[22,52] using the scripts provided in the supplementary materials (available at http://links.lww.com/PAIN/A877). Alternatively, the models can be included directly in statistical analyses through the structural equation modeling framework.

Our results rest on a large population-representative patient sample obtained from a nearly complete database of Sweden's pain specialist treatment clinics, and should therefore be robust and generalizable. The more important limitations to consider include a possible response bias due to varying physical and social settings during data collection, and from the systematic completion of the 3 questionnaires in the same order, which partially explains the higher data attrition for the SF-36. Amounts of missing data were, however, acceptable and sensitivity analyses supported the fact that missingness did not result in any meaningful bias. Another limitation was the observational design, which necessitated the assumption of identical item–trait relationships irrespective of patient characteristics. This assumption should be tested in future measurement invariance studies. Finally, internal consistency estimates should be interpreted with caution because they summarize reliability into a single value, whereas in reality, it varies across the latent trait.

5. Conclusions

This study evaluated the measurement properties of SF-36, EQ-5D, and HADS for chronic pain patients in clinical settings. Our results support that SF-36 is an acceptable measure of 2 independent constructs of physical and mental health. In contrast, although it was a valid approach to summarize the HRQoL construct of EQ-5D as a unidimensional score, its low reliability rendered practical model implementation of dubious value. Finally, rather than dividing into 2 subscales of anxiety and depression, HADS was a valid and reliable measure of overall emotional distress. Relationships between the measured constructs were consistent with their underlying theories, which further supports their construct validity. We recommend that the provided models be used to score SF-36 and HADS in chronic pain patients.

Conflict of interest statement

The authors have no conflicts of interest to declare.

Acknowledgements

The authors thank Lea Constan for proofreading the manuscript, Paolo Frumento for statistical advice, and Phil Chalmers for the excellent statistical open-source package "mirt". The authors also thank the patients, the specialist treatment clinics, and the Swedish Quality Registry of Pain Rehabilitation for their efforts in making this study possible.
This study was funded by the Swedish Research Council (Vetenskapsrådet: 2015-02512) and by the Swedish Research Council for Health, Working Life and Welfare (FORTE: 2017-00177).

Appendix A. Supplemental digital content

Supplemental digital content associated with this article can be found online at http://links.lww.com/PAIN/A877.

Article history:
Received 7 March 2019
Received in revised form 26 July 2019
Accepted 6 August 2019
Available online 13 September 2019

References

[1] Beals J, Welty TK, Mitchell CM, Rhoades DA, Yeh JL, Henderson JA, Manson SM, Buchwald DS. Different factor loadings for SF36: the Strong Heart Study and the National Survey of Functional Health Status. J Clin Epidemiol 2006;59:208–15.
[2] Bjelland I, Dahl AA, Haug TT, Neckelmann D. The validity of the Hospital Anxiety and Depression Scale. An updated literature review. J Psychosom Res 2002;52:69–77.
[3] Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika 1981;46:443–59.
[4] Borsboom D, Mellenbergh GJ, van Heerden J. The theoretical status of latent variables. Psychol Rev 2003;110:203–19.
[5] Brandberg Y, Bolund C, Sigurdardottir V, Sjödén PO, Sullivan M. Anxiety and depressive symptoms at different stages of malignant melanoma. Psycho Oncol 1992;1:71–8.
[6] Brazier J, Connell J, Papaioannou D, Mukuria C, Mulhern B, Peasgood T, Jones ML, Paisley S, O'Cathain A, Barkham M, Knapp M, Byford S, Gilbody S, Parry G. A systematic review, psychometric analysis and qualitative assessment of generic preference-based measures of health in mental health populations and the estimation of mapping functions from widely used specific measures. Health Technol Assess 2014;18:vii–viii, xiii–xxv, 1–188.
[7] Brooks R. EuroQol: the current state of play. Health Policy 1996;37:53–72.
[8] Brooks R. The EuroQol group after 25 years. Dordrecht: Springer, 2013.
[9] Bushnell DM, Reilly MC, Galani C, Martin ML, Ricci JF, Patrick DL, McBurney CR. Validation of electronic data capture of the irritable bowel syndrome—quality of life measure, the work productivity and activity impairment questionnaire for irritable bowel syndrome and the EuroQol. Value Health 2006;9:98–105.
[10] Caci H, Bayle FJ, Mattei V, Dossios C, Robert P, Boyer P. How does the Hospital and Anxiety and Depression Scale measure anxiety and depression in healthy subjects? Psychiatry Res 2003;118:89–99.
[11] Cai L. Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. J Educ Behav Stats 2010:307–35.

[12] Cai L. A two-tier full-information item factor analysis model with applications. Psychometrika Monogr 2010;75:581–612.
[13] Cai L, Hansen M. Limited-information goodness-of-fit testing of hierarchical item factor models. Br J Math Stat Psychol 2013;66:245–76.
[14] Chalmers RP. mirt: a multidimensional item response theory package for the R environment. J Stat Softw 2012;48:1–29.
[15] Chiarotto A, Boers M, Deyo RA, Buchbinder R, Corbin TP, Costa LOP, Foster NE, Grotle M, Koes BW, Kovacs FM, Lin CC, Maher CG, Pearson AM, Peul WC, Schoene ML, Turk DC, van Tulder MW, Terwee CB, Ostelo RW. Core outcome measurement instruments for clinical trials in nonspecific low back pain. PAIN 2018;159:481–95.
[16] Chiarotto A, Deyo RA, Terwee CB, Boers M, Buchbinder R, Corbin TP, Costa LO, Foster NE, Grotle M, Koes BW, Kovacs FM, Lin CW, Maher CG, Pearson AM, Peul WC, Schoene ML, Turk DC, van Tulder MW, Ostelo RW. Core outcome domains for clinical trials in non-specific low back pain. Eur Spine J 2015;24:1127–42.
[17] Chiarotto A, Terwee CB, Kamper SJ, Boers M, Ostelo RW. Evidence on the measurement properties of health-related quality of life instruments is largely missing in patients with low back pain: a systematic review. J Clin Epidemiol 2018;102:23–37.
[18] Coons SJ, Rao S, Keininger DL, Hays RD. A comparative review of generic quality-of-life instruments. Pharmacoeconomics 2000;17: 13–35.
[19] Cosco TD, Doyle F, Ward M, McGee H. Latent structure of the hospital anxiety and depression scale: a 10-year systematic review. J Psychosom Res 2012;72:180–4.
[20] Coyne JC, van Sonderen E. No further research needed: abandoning the Hospital and Anxiety Depression Scale (HADS). J Psychosom Res 2012; 72:173–4.
[21] de Vet HC, Ader HJ, Terwee CB, Pouwer F. Are factor analytical techniques used appropriately in the validation of health status questionnaires? A systematic review on the quality of factor analysis of the SF-36. Qual Life Res 2005;14:1203–18; discussion 1219–1221, 1223–1204.
[22] DeMars C. A tutorial on interpreting bifactor model scores. Int J Test 2013;13:354–78.
[23] Dunbar M, Ford G, Hunt K, Der G. A confirmatory factor analysis of the Hospital Anxiety and Depression scale: comparing empirically and theoretically derived structures. Br J Clin Psychol 2000;39:79–94.
[24] Edwards RR, Dworkin RH, Turk DC, Angst MS, Dionne R, Freeman R, Hansson P, Haroutounian S, Arendt-Nielsen L, Attal N, Baron R, Brell J, Bujanover S, Burke LB, Carr D, Chappell AS, Cowan P, Etropolski M, Fillingim RB, Gewandter JS, Katz NP, Kopecky EA, Markman JD, Nomikos G, Porter L, Rappaport BA, Rice AS, Scavone JM, Scholz J, Simon LS, Smith SM, Tobias J, Tockarshewsky T, Veasley C, Versavel M, Wasan AD, Wen W, Yarnitsky D. Patient phenotyping in clinical trials of chronic pain treatments: IMMPACT recommendations. PAIN 2016;157: 1851–71.
[25] EuroQol G. EuroQol—a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199–208.
[26] Farivar SS, Cunningham WE, Hays RD. Correlated physical and mental health summary scores for the SF-36 and SF-12 Health Survey, V.I. Health Qual Life Outcomes 2007;5:54.
[27] Feng YS, Jiang R, Kohlmann T, Pickard AS. Exploring the internal structure of the EQ-5D using non-preference-based methods. Value Health 2019;22:527–36.
[28] Friedman S, Samuelian JC, Lancrenon S, Even C, Chiarelli P. Three-dimensional structure of the Hospital Anxiety and Depression Scale in a large French primary care population suffering from major depression. Psychiatry Res 2001;104:247–57.
[29] Gamst-Klaussen T, Gudex C, Olsen JA. Exploring the causal and effect nature of EQ-5D dimensions: an application of confirmatory tetrad analysis and confirmatory factor analysis. Health Qual Life Outcomes 2018;16:153.
[30] Gibbons RD, Darrell RB, Hedeker D, Weiss DJ, Segawa E, Bhaumik DK, Kupfer DJ, Frank E, Grochocinski VJ, Stover A. Full-information item bifactor analysis of graded response data. Appl Psychol Meas 2007;31: 4–19.
[31] Grassi M, Nucera A; European Community Respiratory Health Study Quality of Life Working G. Dimensionality and summary measures of the SF-36 v1.6: comparison of scale- and item-based approach across ECRHS II adults population. Value Health 2010;13:469–78.
[32] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer, 2017.
[33] Hays RD, Sherbourne CD, Mazel RM. The RAND 36-item health survey 1.0. Health Econ 1993;2:217–27.
[34] Johnsen LG, Hellum C, Nygaard OP, Storheim K, Brox JI, Rossvoll I, Leivseth G, Grotle M. Comparison of the SF6D, the EQ5D, and the oswestry disability index in patients with chronic low back pain and degenerative disc disease. BMC Musculoskelet Disord 2013;14:148.
[35] Kaiser U, Kopkow C, Deckert S, Neustadt K, Jacobi L, Cameron P, De Angelis V, Apfelbacher C, Arnold B, Birch J, Bjarnegard A, Christiansen S, C de C Williams A, Gossrau G, Heinks A, Huppe M, Kiers H, Kleinert U, Martelletti P, McCracken L, de Meij N, Nagel B, Nijs J, Norda H, Singh JA, Spengler E, Terwee CB, Tugwell P, Vlaeyen JWS, Wandrey H, Neugebauer E, Sabatowski R, Schmitt J. Developing a core outcome domain set to assessing effectiveness of interdisciplinary multimodal pain therapy: the VAPAIN consensus statement on core outcome domains. PAIN 2018;159:673–83.
[36] Kang T, Chen TT. Performance of the generalized S-X^2 item fit index for polytomous IRT models. J Educ Meas 2007;45:391–406.
[37] Karimi M, Brazier J. Health, health-related quality of life, and quality of life: what is the difference? Pharmacoeconomics 2016;34:645–9.
[38] Maydeu-Olivares A. Further empirical results on parametric versus non-parametric IRT modeling of Likert-type personality data. Multivariate Behav Res 2005;40:261–79.
[39] Maydeu-Olivares A. Goodness-of-fit assessment of item response theory models. Measurement 2013;11:71–101.
[40] Moorey S, Greer S, Watson M, Gorman C, Rowden L, Tunmore R, Robertson B, Bliss J. The factor structure and factor stability of the hospital anxiety and depression scale in patients with cancer. Br J Psychiatry 1991;158:255–9.
[41] Nicholas M, Vlaeyen JWS, Rief W, Barke A, Aziz Q, Benoliel R, Cohen M, Evers S, Giamberardino MA, Goebel A, Korwisi B, Perrot S, Svensson P, Wang SJ, Treede RD; IASP Taskforce for the Classification of Chronic Pain. The IASP classification of chronic pain for ICD-11: chronic primary pain. PAIN 2019;160:28–37.
[42] Norton S, Cosco T, Doyle F, Done J, Sacker A. The Hospital Anxiety and Depression Scale: a meta confirmatory factor analysis. J Psychosom Res 2013;74:74–81.
[43] Nyberg V, Sanne H, Sjolund BH. Swedish quality registry for pain rehabilitation: purpose, design, implementation and characteristics of referred patients. J Rehabil Med 2011;43:50–7.
[44] Pallant JF, Bailey CM. Assessment of the structure of the hospital anxiety and depression scale in musculoskeletal patients. Health Qual Life Outcomes 2005;3:82.
[45] Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). Br J Clin Psychol 2007;46:1–18.
[46] Prieto L, Novick D, Sacristan JA, Edgell ET, Alonso J, Group SS. A Rasch model analysis to test the cross-cultural validity of the EuroQoL-5D in the schizophrenia outpatient health outcomes study. Acta Psychiatr Scand Suppl 2003;107:24–9.
[47] Razavi D, Delvaux N, Farvacques C, Robaye E. Screening for adjustment disorders and major depressive disorders in cancer in-patients. Br J Psychiatry 1990;156:79–83.
[48] Reckase M. Multidimensional item response theory. New York: Springer, 2009.
[49] Reed PJ. Medical outcomes study short form 36: testing and cross-validating a second-order factorial structure for health system employees. Health Serv Res 1998;33:1361–80.
[50] Reise SP. A comparison of item- and person-fit methods of assessing model-data fit in IRT. Appl Psychol Meas 1990;14:127–37.
[51] Reise SP, Moore TM, Haviland MG. Bifactor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. J Personal Assess 2010;92:544–59.
[52] Reise SP, Revicki DA. Handbook of item response theory modeling: applications to typical performance assessment. New York: Routledge, 2015.
[53] Rijmen F. Formal relations and an empirical comparison among the Bi-factor, the testlet, and a second-order multidimensional IRT model. J Educ Meas 2010;47:361–72.
[54] Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monogr 1969;34:1–97.
[55] Steingrimsdottir OA, Landmark T, Macfarlane GJ, Nielsen CS. Defining chronic pain in epidemiological studies: a systematic review and meta-analysis. PAIN 2017;158:2092–107.
[56] Stochl J, Croudace T, Perez J, Birchwood M, Lester H, Marshall M, Amos T, Sharma V, Fowler D, Jones PB, National Eden Study T. Usefulness of EQ-5D for evaluation of health-related quality of life in young adults with first-episode psychosis. Qual Life Res 2013;22:1055–63.
[57] Streiner D, Norman G, Cairney J. Health measurement scales: a practical guide to their development and use. Oxford: Oxford University Press, 2015.
[58] Sullivan M, Karlsson J, Ware JE Jr. The Swedish SF-36 Health Survey—I. Evaluation of data quality, scaling assumptions, reliability and construct

validity across general populations in Sweden. Soc Sci Med 1995;41: 1349–58.
[59] Taft C, Karlsson J, Sullivan M. Do SF-36 summary component scores accurately summarize subscale scores? Qual Life Res 2001;10:395–404.
[60] Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60: 34–42.
[61] Toye F, Seers K, Hannink E, Barker K. A mega-ethnography of eleven qualitative evidence syntheses exploring the experience of living with chronic non-malignant pain. BMC Med Res Methodol 2017;17:116.
[62] Treede RD, Rief W, Barke A, Aziz Q, Bennett MI, Benoliel R, Cohen M, Evers S, Finnerup NB, First MB, Giamberardino MA, Kaasa S, Korwisi B, Kosek E, Lavand'homme P, Nicholas M, Perrot S, Scholz J, Schug S, Smith BH, Svensson P, Vlaeyen JWS, Wang SJ. Chronic pain as a symptom or a disease: the IASP classification of chronic pain for the international classification of diseases (ICD-11). PAIN 2019;160:19–27.
[63] Tucker G, Adams R, Wilson D. Observed agreement problems between sub-scales and summary components of the SF-36 version 2—an alternative scoring method can correct the problem. PLoS One 2013;8: e61191.
[64] Turk DC, Dworkin RH, Allen RR, Bellamy N, Brandenburg N, Carr DB, Cleeland C, Dionne R, Farrar JT, Galer BS, Hewitt DJ, Jadad AR, Katz NP, Kramer LD, Manning DC, McCormick CG, McDermott MP, McGrath P, Quessy S, Rappaport BA, Robinson JP, Royal MA, Simon L, Stauffer JW, Stein W, Tollett J, Witter J. Core outcome domains for chronic pain clinical trials: IMMPACT recommendations. PAIN 2003;106:337–45.
[65] Ware JE Jr, Kosinski M, Gandek B, Aaronson NK, Apolone G, Bech P, Brazier J, Bullinger M, Kaasa S, Leplege A, Prieto L, Sullivan M. The factor structure of the SF-36 health survey in 10 countries: results from the IQOLA project. International Quality of Life Assessment. J Clin Epidemiol 1998;51:1159–65.
[66] Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473–83.
[67] Ware JE, Snow KK, Kosinski M, Gandek B. SF-36 health survey: manual and interpretation guide. Lincoln: Quality Metric, 1993.
[68] Wilson D, Parsons J, Tucker G. The SF-36 summary scales: problems and solutions. Soz Praventivmed 2000;45:239–46.
[69] Wirth RJ, Edwards MC. Item factor analysis: current approaches and future directions. Psychol Methods 2007;12:58–79.
[70] Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983;67:361–70.