Osteoporos Int (2006) 17: 543–551
DOI 10.1007/s00198-005-0024-7

ORIGINAL ARTICLE

Development of the Qualeffo–31, an osteoporosis-specific quality-of-life questionnaire

N. M. van Schoor · D. L. Knol · C. A. W. Glas · R. W. J. G. Ostelo · A. Leplège · C. Cooper · O. Johnell · P. Lips

Received: 18 May 2005 / Accepted: 3 October 2005 / Published online: 14 December 2005

Abstract *Introduction:* Vertebral deformities are a common consequence of osteoporosis and are known to decrease quality of life. The Qualeffo–41 is a quality-of-life questionnaire especially developed for measuring quality of life in patients with vertebral deformities. It consists of 41 questions arranged in five domains: pain, physical function, social function, general health perception, and mental function. The objectives of this study were: (1) to develop a shorter version of the Qualeffo–41 by removing redundant questions; and (2) to investigate the scale characteristics, reliability, and validity of this shorter version. *Methods:* The study was performed using data from the Qualeffo validation study and the Multiple Outcomes of Raloxifene Evaluation (MORE) study. The analyses were performed in patients with vertebral deformities (n=579). Factor analysis on polychoric correlations and an item response theory (IRT) model, i.e., the generalized partial credit model (GPCM), were used to create a shorter version of Qualeffo–41. Using GPCM, scoring weights were computed for all items. *Results:* Three items were removed from the data set because of too many missing values. Factor analysis identified three instead of five domains: (1) pain, (2) physical function, and (3) mental function. Five items had factor loadings <0.4 and were not included in the GPCM. After excluding several items, the domains pain (four items), physical function (18 items), and mental function (nine items) showed a good, reasonable, and excellent fit, respectively. This indicates that the mental function domain and the pain domain are more unidimensional than the physical function domain. All three domains showed a very high correlation ($r \geq 0.95$) with the corresponding domains of the Qualeffo–41. *Conclusions:* Qualeffo–31 was developed, consisting of three domains with a reasonable to excellent fit to the GPCM. Although the fit to the GPCM supports the construct validity of the Qualeffo–31, validation in a new study should be performed before using it in practice.

Keywords Item response theory · Qualeffo–31 · Qualeffo–41 · Quality of life · Vertebral fracture

N. M. van Schoor · R. W. J. G. Ostelo · P. Lips
Institute for Research in Extramural Medicine,
VU University Medical Center,
Amsterdam, The Netherlands

D. L. Knol
Department of Clinical Epidemiology and Biostatistics,
VU University Medical Center,
Amsterdam, The Netherlands

C. A. W. Glas
Department of Educational Measurement and Data Analysis,
University of Twente,
Enschede, The Netherlands

A. Leplège
Institut National de la Santé et de la
Recherche Médicale (INSERM),
Unité 292, Hôpital de Bicêtre,
Le Kremlin–Bicêtre, France

C. Cooper
MRC Epidemiology Resource Centre,
University of Southampton,
Southampton, UK

O. Johnell
Department of Orthopaedics, Malmö General Hospital,
Malmö, Sweden

P. Lips (✉)
Department of Endocrinology, VU University Medical Center,
P.O. Box 7057, 1007 Amsterdam, The Netherlands
e-mail: p.lips@vumc.nl
Tel.: +31-20-4440614
Fax: +31-20-4440502

Introduction

A common consequence of osteoporosis is the occurrence of vertebral deformities. About 25% of the elderly population suffers from vertebral deformities [1–6]. Vertebral deformities may cause pain, decreased physical functioning, social isolation, and depression [3, 5–8]. These aspects can be captured by the concept of quality of life.

Quality of life can be measured by disease-specific instruments, which are targeted to one disease or a group of diseases, and generic instruments, which are developed for general use. Disadvantages of generic instruments is that they usually contain superfluous questions and that these questionnaires are less responsive to change in specific subpopulations. Therefore, especially in randomized controlled trials, it can be an advantage to use a disease-specific questionnaire.

An example of a quality-of-life questionnaire specifically developed for persons with vertebral deformities is the Qualeffo–41. This questionnaire was developed in 1996 by the Working Party of the European Foundation for Osteoporosis [7, 9, 10]. Originally, the questionnaire consisted of 48 questions and six visual analogue scales. In the Qualeffo validation study, the number of items was reduced to 41 and the visual analogue scales were removed. This resulted in the Qualeffo–41, which consists of 41 questions in five domains: pain, physical function, social function, general health perception, and mental function. A disadvantage of the Qualeffo–41 is the large number of items. As described before, Qualeffo was originally developed by experts in the field. As a next step, the current study aims to develop a shorter and more practical instrument by using the item response theory (IRT). Traditionally, the classical test theory (CTT) is often used to evaluate questionnaires. Using CTT, a person's (dis)ability (here, quality of life) is described by summing all items. A more modern approach is IRT, which describes the association between a respondent's underlying level on a latent trait and the probability of a particular item response [11]. An advantage of IRT as compared with CTT is that separately from a person's (dis)ability, difficulty parameters of the items can be estimated [11]. In addition, in the IRT model used in this study, the generalized partial credit model (GPCM), item discrimination parameters are computed [11, 12]. Higher values of this parameter are associated with items that are better able to discriminate between contiguous trait levels [11]. These item discrimination parameters or scoring weights are used in calculating the sum score of a domain, yielding a more precise estimate. A second aim of this study is to investigate the scale characteristics, reliability, and validity of the shorter Qualeffo version in our own data set.

Materials and methods

Design and subjects

The analyses were performed using data from the Qualeffo validation study (131 persons with vertebral fractures; 182 without) [7] and baseline data from the Multiple Outcomes of Raloxifene Evaluation (MORE) study (448 persons with vertebral fractures; 301 without) [13]. From both studies, only those subjects were included who had vertebral fractures (n=579) because Qualeffo–41 was specifically developed for patients with vertebral fractures [9]. The Qualeffo validation study was performed in seven different European centers (Liège, Paris, Bad Pyrmont, Cambridge, Siena, Amsterdam, and Malmö), and the MORE data used in this study were obtained in seven European countries (Belgium, France, Germany, Great Britain, Italy, The Netherlands, and Sweden). Both studies were performed in osteoporotic patients.

Preparation of the data set

First, the answers of the Qualeffo–41 were, when necessary, transformed so that the lowest score represents the best and the highest score represents the worst quality of life [7]. Second, all answers were recoded so that all items had a range from 1 to 5 [7]. Third, items with more than 25% missing values were removed from the questionnaire, and cases with more than 50% missing values were removed from the data set.

Assessing the domains by factor analysis

Because of the five-point ordinal scale of the items, a polychoric correlation matrix was made using PRELIS 2.5.4 [14]. A polychoric correlation measures the linear relationship between two observed, discrete variables that are manifestations of latent, normal, continuous variables. Therefore, a polychoric correlation is considered a more appropriate measure of the relationship between two Likert-type items than the Pearson correlation [15]. A factor analysis was performed in SPSS 11.0.1 to see whether the results were different when using a polychoric correlation matrix as compared with the Pearson correlation matrix. In a second sensitivity analysis, the missing values were replaced according to the Expectation Maximization (EM) algorithm in SPSS. This was compared with a factor analysis using pairwise deletion of missing data.

Because we did not know whether the same domains would be important in the shorter version of the Qualeffo–41, we used exploratory factor analysis, and we did not fix the number of factors. Three criteria were used to determine the number of factors: (1) the eigenvalue should be larger than one; (2) the scree test; and (3) the content of the factors. Roughly, the eigenvalue is a measure of the variance accounted for by that factor and can be thought of as the contribution of the factor [16]. If the eigenvalue is larger than one for a factor, it explains at least as much variance as a single variable. It is the most common criterion but often results in too many factors. Therefore, as a second criterion, the scree test was used—a graph of the eigenvalues for each factor in which one looks for a break in the distribution [16]. Items with an absolute factor loading less than 0.4 in the pattern matrix of the principal component analysis with Promax rotation (default) were excluded.

Assessing the fit of the domains by item response theory

A unidimensional IRT model, the GPCM [17], was fitted to each of the three dimensions identified using factor anal-

ysis. Poorly fitting items were removed using two closely related criteria. To compute the criteria, the score range was divided into four score levels, in such a way that the numbers of respondents in these score groups were approximately equal. Using this partition of the sample, a Lagrange Multiplier (LM) test was computed for every item to test whether the observed and expected item score in the score groups matched (for details [18, 19]). Items with a value on the LM statistic larger than 30 were removed from the questionnaire. The remaining item pool still had a lot of items with significant values on the LM test. This is explained by the power of the test: with larger sample sizes, differences between observed and expected item scores that have little practical implications become statistically significant. Therefore, as a second criterion, items with a mean absolute difference between the observed and expected item scores in the four score groups of more than 0.15 were removed. Finally, the estimated discrimination parameters were used as weights for the calculation of the scores of the shorter version of the Qualeffo–41.

Scale characteristics, reliability, and validity

Analyses on scale characteristics, reliability, and validity were largely performed as described in the Qualeffo validation study [7]. In addition, the criteria used were based on this article. First, floor effects, i.e., the percentage of subjects with the lowest possible domain score, and ceiling effects, i.e., the percentage of subjects with the highest

Table 1 Factors in the pattern matrix of the principal component analysis with Promax rotation

Items	Description	Factor		
		Pain	Physical function	Mental function
1	Frequency back pain	.872	–	–
2	Length back pain during daytime	.924	–	–
3	Severity back pain	.834	–	–
4	Back pain at other times	.691	–	–
5	Back pain disturbing sleep	.733	–	–
6	Dressing	.404	.497	–
7	Taking a bath or shower	–	.617	–
8	Getting to or operating a toilet	–	.543	–
9	Sleeping	–	–	–
10	Cleaning	–	.603	–
11	Preparing meals	–	.695	–
12	Washing the dishes	–	.664	–
13	Day-to-day shopping	–	.784	–
14	Lifting a heavy object	–	.643	–
15	Get up from a chair	–	.595	–
16	Bend down	–	.603	–
17	Kneel down	–	.721	–
18	Climb stairs	–	.788	–
19	Walk 100 yards	–	.851	–
20	Being outside	–	.785	–
21	Using public transport	–	.879	–
22	Affected by changes of figure due to osteoporosis	–	–	–
24	Gardening	–	.793	–
26	Visit a cinema, theatre, etc.	–	.865	–
27	Visit friends or relatives	–	.588	–
28	Participate in social activities	−.559	.734	–
30	Health	–	–	–
31	Overall quality of life during the last week	–	–	.468
32	Overall quality of life compared with 10 years ago	–	–	–
33	Feel tired	–	–	.418
34	Feel downhearted	–	–	.857
35	Feel lonely	–	–	.747
36	Feel full of energy	–	–	.441
37	Hopeful about future	–	–	.767
38	Get upset over little things	–	–	.624
39	Easy to make contact	–	–	.460
40	In good spirits	–	–	.787
41	Afraid of becoming totally dependent	–	–	–

Presented are the factor loadings with an absolute value >0.4

possible domain score, were determined. Second, the internal consistency was determined by calculating Cronbach's alpha per domain. Third, the correlation between the domain scores was assessed. Fourth, convergent validity was tested, i.e., the correlation between the score for each question and its own domain score should be higher than 0.40. Fifth, discriminant validity was tested, i.e., the correlation of the score for each question with its own domain score should be higher than with the scores of other domains. Sixth, the domain scores of the Qualeffo–31 were correlated with the domain scores of the Qualeffo–41. Finally, the average Qualeffo–31 and Qualeffo–41 scores were presented for persons without vertebral fractures and for persons with vertebral fractures. In addition, average scores were presented for persons with 1 and ≥2 vertebral fractures. For these analyses, the controls of the Qualeffo validation study and the MORE study were included.

Results

Three items were removed from the questionnaire because of too many missing values. Items 23, 25, and 29 had 207, 181, and 261 missing values, respectively. Item 29, about

Table 2 Final fit to Final fit to generalized partial credit model (GPCM)

Item [a]	Sign. prob.	Mean deviance	Weight	Difficulty (1–2)	Difficulty (2–3)	Difficulty (3–4)	Difficulty (4–5)	Group 1		Group 2		Group 3		Group 4	
								Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.
For domain "Pain"															
1(1)	.00	.10	2.785	−0.76	−0.41	0.63	−0.06	1.41	1.34	2.90	3.02	3.49	3.57	4.00	3.96
2(2)	.02	.09	2.858	−0.93	0.38	0.69	0.42	0.75	0.72	2.10	2.08	2.73	2.80	3.48	3.60
3(3+4)[b]	.14	.06	1.286	−0.46	0.10	2.33	2.25	0.59	0.61	1.28	1.31	1.79	1.70	2.12	2.12
4(5)	.02	.08	0.686	2.66	0.15	1.53	−0.38	0.16	0.22	0.75	0.73	1.50	1.37	2.42	2.49
For domain "Physical function"															
1(6)	.79	.01	1.671	1.06	1.71	1.98	3.45	0.03	0.03	0.11	0.12	0.37	0.34	1.15	1.17
2(7)	.00	.04	1.749	0.95	1.63	1.70	1.81	0.01	0.03	0.15	0.13	0.44	0.39	1.49	1.53
3(8)	.73	.02	1.441	2.11	1.87	3.20	2.64	0.01	0.01	0.06	0.04	0.10	0.11	0.55	0.55
4(10)	.09	.06	1.663	−0.45	0.63	1.63	1.80	0.26	0.31	0.94	0.85	1.43	1.41	2.34	2.40
5(11)	.31	.04	1.676	0.94	1.65	1.97	2.52	0.03	0.04	0.17	0.15	0.34	0.39	1.33	1.30
6(12)	.74	.02	1.819	1.01	1.48	2.22	3.29	0.02	0.03	0.11	0.12	0.37	0.35	1.21	1.22
7(13)	.01	.04	1.871	0.21	1.07	1.41	1.61	0.05	0.10	0.43	0.40	0.98	0.94	2.29	2.31
8(14)	.00	.10	1.008	−1.13	−0.03	0.02	0.13	1.06	1.01	2.25	2.20	2.82	3.02	3.68	3.61
9(15)	.02	.05	1.485	0.59	1.70	2.51	2.74	0.04	0.08	0.28	0.27	0.59	0.55	1.27	1.28
10(16)	.01	.07	1.407	−0.34	0.47	1.83	2.45	0.26	0.32	1.00	0.88	1.37	1.41	2.21	2.21
11(17)	.13	.08	1.072	−0.63	0.45	1.42	1.15	0.47	0.51	1.22	1.13	1.77	1.73	2.64	2.72
12(18)	.05	.04	1.640	0.21	1.35	2.28	1.82	0.09	0.12	0.39	0.40	0.84	0.79	1.70	1.71
13(19)	.09	.03	1.520	0.18	1.41	2.24	2.01	0.12	0.14	0.40	0.42	0.82	0.80	1.69	1.67
14(20)	.09	.06	0.488	4.04	−0.26	1.83	2.08	0.22	0.21	0.55	0.45	0.71	0.79	1.56	1.57
15(21)	.05	.06	2.130	0.70	1.02	1.55	1.16	0.02	0.03	0.23	0.18	0.58	0.66	2.40	2.34
16(24)	.46	.04	1.981	−0.76	0.93	–	–	0.34	0.36	0.90	0.85	1.19	1.21	1.64	1.65
17(26)	.07	.06	1.572	1.01	1.12	–	–	0.07	0.04	0.16	0.17	0.38	0.46	1.26	1.19
18(27)	.01	.09	0.455	1.35	2.61	2.64	–	0.32	0.34	0.68	0.53	0.59	0.72	1.09	1.08
For domain "Mental function"															
1(31)	.62	.04	1.062	−1.90	−0.06	0.97	2.64	0.99	1.03	1.52	1.51	1.93	1.91	2.51	2.51
2(33)	.10	.06	0.467	−2.35	−0.13	0.57	1.39	1.51	1.44	1.88	1.87	2.13	2.24	2.73	2.72
3(34)	.04	.05	1.906	−0.39	1.43	1.59	0.77	0.25	0.24	0.63	0.56	0.91	0.92	1.99	2.05
4(35)	.08	.07	0.891	0.79	2.68	1.33	0.02	0.13	0.17	0.48	0.34	0.58	0.60	1.55	1.59
5(36)	.89	.03	0.679	0.53	0.99	−1.35	1.98	0.62	0.64	1.27	1.26	1.97	1.97	2.86	2.84
6(37)	.50	.03	0.997	−0.59	−0.05	1.32	2.10	0.68	0.64	1.13	1.15	1.65	1.65	2.38	2.41
7(38)	.07	.04	0.544	−1.26	−0.59	2.39	4.77	1.09	1.04	1.38	1.39	1.66	1.65	1.97	2.03
8(39)	.49	.03	0.355	0.80	2.11	3.64	2.05	0.53	0.56	0.75	0.75	0.94	0.95	1.38	1.34
9(40)	.69	.03	1.610	−0.70	1.04	1.97	2.60	0.41	0.38	0.68	0.72	1.06	1.06	1.74	1.73

"Sign. prob." gives the significance probability of the Lagrange Multiplier (LM) test; "Mean deviance" gives the mean difference between the observed and expected item scores in the four score groups; "Weight" gives the estimates of the item discrimination parameters; "Difficulty" gives the level of disability at which 50% of the people will respond in a higher answer category of two adjacent answer categories (1–2 means from answer category 1 to 2); Obs. and Exp. give the observed and expected item scores in the four score groups

[a] Original item number Qualeffo–41 between brackets

[b] Items 3 and 4 of the original questionnaire were combined (see also Discussion)

intimacy, had a "not applicable" option. Because the non-applicable option was entered as a system missing value in both data sets, we were not able to identify which persons did not have intimacy and which persons did not answer this question for other reasons. In addition, three cases were removed from the analyses because of more than 50% missing values. The remaining missing values were considered as missing at random (MAR), except for items 24 (about gardening) and 26 (about visiting a cinema or theater), which were the only remaining items with a nonapplicable option.

In the sensitivity analyses, a polychoric correlation matrix was compared with a Pearson correlation matrix. When using a polychoric correlation matrix, six factors were identified as defined by the eigenvalue >1 criterion, and the explained variance was 64%; when using the Pearson correlation matrix, seven factors were identified and the explained variance was 58%. Replacement of the missing values did not influence the results. This indicates that items 24 and 26, which had missing values not at random (MNAR), did not influence the results. While fewer factors were identified with higher explained variance using a polychoric correlation matrix and replacement of missing values had no influence, we decided to perform further analyses using a polychoric correlation matrix with pairwise deletion of missing data.

In the factor analyses, six factors were identified using the criterion that the eigenvalue should be larger than one, and two or three factors when using the scree plot. When using six, five, or four domains, there were two, two, and one factor(s), respectively, that had three items or less. Beside the fact that these domains were very small, they seemed meaningless. Therefore, it was decided to start further analyses with two or three domains. In the model containing two domains, a physical function and a mental function domain were identified (data not shown). In the model containing three domains, a pain domain, a physical function domain, and a mental function domain were identified (Table 1). The advantage of the questionnaire with three domains is that the pain domain, which contains specific questions for patients with vertebral fractures, is retained. While the Qualeffo is a disease-specific questionnaire, we chose to continue with the three-domain model. In this model, items 9, 22, 30, 32, and 41 had a factor loading <0.4 and were not included in the GPCM. In addition, items 6 and 28, which loaded both on the pain domain and the physical function domain, were only included in the physical function domain because of the content of the items. Another reason is that item 28 acted as a suppressor in the pain domain.

Using GPCM, all items with an LM statistic larger than 30 and items where the mean absolute difference between the observed and expected item scores in the four score groups was larger than 0.15 were removed. In Table 2 the final domains are presented. Because the pain model contains very few questions, item 1 with an LM statistic of 32.03 was also retained. Note that both the observed and expected item scores increase over the score groups. In addition, the difficulties of the answer categories, i.e., the level of disability at which 50% of the people will respond in the higher category of two adjacent answer categories [20], increases for most items. In Figs. 1 and 2, examples of category curves of two different items are presented. The difficulties are the intersections between the answer categories. The first item of the mental function domain is an example of a good item while the categories intersect in the good order (category 1 with 2, 2 with 3, 3 with 4, 4 with 5). This means that a higher level of disability (or a lower quality of life during the past week) is associated with an increased probability of scoring a higher answer category. In contrast, in the first item of the pain domain, categories 2 and 3 are scored more often than category 4 for all levels of disability. In other words, according to the model, respondents prefer categories 2 or 3 to 4, irrespective their state of disability [20].

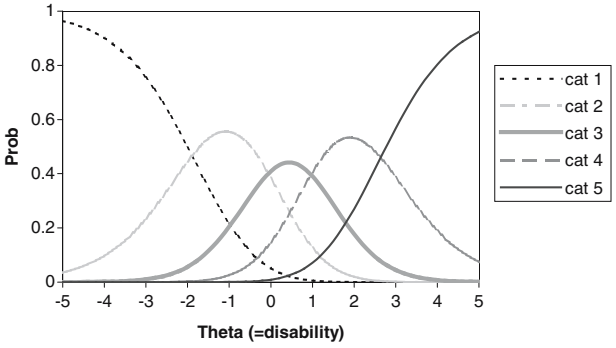

Fig. 1 Category curves of first item of mental function domain

After calculating the domain scores of the Qualeffo–31, it was seen that the distributions of the domains of the Qualeffo–31 were very similar to those of the Qualeffo–41: the original and new physical function domain showed a distribution skewed to the right, and the original and new mental function domains were normally distributed. Also, the new pain domain looked very similar to the original pain domain. However, here a floor-effect was present. For the pain domain, 77 persons (13.6%) had the lowest score; for the physical function domain, 28 persons (4.9%) had the lowest score; and for the mental function domain, five persons had the lowest score (0.9%). None of the domains showed a ceiling effect.

Cronbach's alpha, a measure for the internal consistency, was 0.72 for the pain domain, 0.93 for the physical function domain, and 0.79 for the mental function domain. The

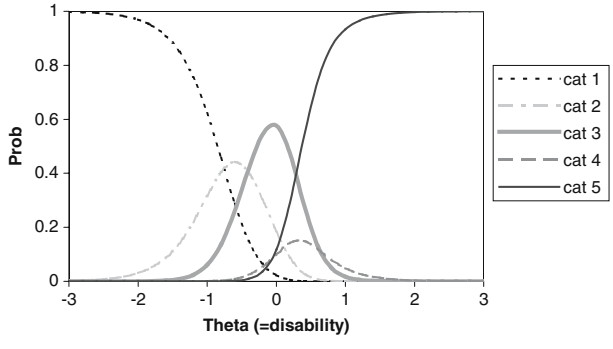

Fig. 2 Category curves of first item of pain domain

Table 3 Comparison of Qualeffo–41 and Qualeffo–31 scores

	Pain–41	Pain–31	Physical function–41	Physical function–31	Mental function–41	Mental function–31
VFX						
No (n=483)	25.7±25.2	23.4±23.8	13.6±12.2	9.8±10.8	29.6±15.4	21.8±14.2
Yes (n=579)	40.1±26.4*	37.1±25.2*	22.6±17.5*	17.3±15.8*	34.6±17.3*	27.1±16.3*
1 VFX (n=240)	36.4±25.7	32.9±24.4	18.7±15.0	14.0±13.5	32.6±16.2	24.9±14.9
≥2 VFX (n=339)	42.7±26.6*	40.0±25.4	25.3±18.6*	19.7±16.9*	36.0±17.9*	28.6±17.0*

Presented are the average scores±standard deviations
VFX=vertebral fracture
*p<0.001 for comparisons "no VFX versus VFX" and "no VFX versus 1 VFX versus ≥2 VFX"

correlation between the new pain domain and the new physical function domain was 0.61, between the new pain domain and the new mental function domain 0.32, and between the new physical function domain and the new mental function domain 0.54.

Convergent validity, i.e., the correlation of each question with its domain score, was larger than 0.40 for all items (100%) of the pain domain, for 17 out of 18 items (94%) of the physical function domain, and for eight out of nine items (89%) of the mental function domain. In addition, discriminant validity, i.e., the correlation of every item (100%) with its domain score, was higher than with the other domain scores. Finally, very high correlations were found between the new and original pain domain (r=0.97), between the new and original physical function domain (r=0.98), and between the new and original mental function domain (r=0.95).

In Table 3, the average Qualeffo–41 and Qualeffo–31 scores are presented for patients with vertebral fractures and for controls. In addition, average scores are presented for persons with 1 and ≥2 vertebral fractures. When comparing the scores, it should be kept in mind that the domain structure from the Qualeffo–41 and Qualeffo–31 differ and that the new domains not only have fewer items but also items from other Qualeffo–41 domains. As can be seen, the scores on the Qualeffo–31 are somewhat lower than the scores on Qualeffo–41. However, the distances between the scores for cases and controls and for persons with 0, 1, ≥2 vertebral fractures seem very similar, and all differences were statistically significant at p<0.001.

Discussion

In this study, the Qualeffo–31, a shorter version of the Qualeffo–41, was developed. Three domains were identified: a pain domain, a physical function domain, and a mental function domain (see Appendix). Looking at the LM statistics and the mean absolute differences in observed and expected item scores, it can be concluded that the pain domain showed a good fit, the physical function domain a reasonable fit, and the mental function domain an excellent fit. This indicates that the mental function domain and the pain domain are more unidimensional than the physical function domain and that in the latter, more aspects are being measured than physical function alone. In addition, the difficulties of the answer categories increase for most items, indicating that a higher level of disability is associated with scoring a higher answer category and that respondents can distinguish correctly between most answer categories. For items that do not have increasing difficulties, future studies should point out whether this is a remaining problem or not although it could indicate that reformulating or reducing the number of answer categories can improve these items. Finally, the internal consistency and convergent and discriminant validity of the domains show good results, and the correlation with the Qualeffo–41 is very high.

A major strength of this study is that factor analysis and GPCM were used to further develop the expert-based Qualeffo. Redundant items were removed from the questionnaire, and the conceptual structure was improved. Factor analysis showed that the items could be rearranged into three domains. The items of two domains, i.e., general health perception and social function, were partly shifted to other domains. General health perception originally consisted of three items, of which one (item 31) about quality of life during the past week was retained and showed a good fit with the mental function domain, which asks about the respondent's mood during the past week. Of the social function domain, items 24, 26, and 27 about gardening, visiting a cinema or theater, and visiting friends or relatives were rearranged to the physical function domain. This is supported by the fact that the content of these items have both a social and a physical component. However, validation of this new conceptual structure of the Qualeffo–31 should be performed in new studies before using the questionnaire in practice, e.g., by examining the fit of the domains in new data sets.

Important advantages for researchers and patients are the lower number of items and the decreased time (and burden) needed to answer the questionnaire. As a consequence, it is easier to combine this questionnaire with other (quality-of-life) questionnaires. As mentioned in the introduction, a disease-specific questionnaire may be more responsive to change. However, in case of a cost-effectiveness analysis, one might want to add a generic questionnaire, such as the EQ–5D, in order to generate utility scores. A disadvantage of Qualeffo–31 as compared with Qualeffo–41 is that it is somewhat more difficult to calculate the domain scores due to the use of weights. However, a SPSS syntax is available, and more precise estimates of quality of life can be made by

using weights. In addition, while we only removed non-fitting items, the questionnaire still has 31 items. As was done with the SF–36, it might be interesting to make a short version in the future, like the SF–12. Finally, it should be stated that when using a questionnaire in an older person, it should be checked whether he/she has sufficient cognitive capabilities to understand and correctly interpret the questions.

Another important issue is that when factor analysis identifies more than one factor, a total score should not be calculated because the questionnaire does not comprise a single construct. As a consequence, the total score is not very informative. For example, if the total quality-of-life score decreases with 9 points for a certain person, it could be that he scores 3 points lower on pain, 3 points lower on physical function, and 3 points lower on mental function. Or, for example, he scores 9 points lower on the physical function domain. However, when a total score is needed for describing the quality of life of a group, e.g., as an outcome in an effectiveness study of osteoporosis medication, an SPSS syntax is available. A limitation of GPCM is that it is not clear which threshold can best be chosen for the LM statistic. Another problem we encountered is that one out of 41 questions of the Qualeffo–41 was not independent (item 4). According to the GPCM, item 3 ("How severe is your back pain at its worst?") was negatively contributing to the fit of the pain domain and was removed. Therefore, item 4 ("How is your back pain at other times?") had to be adapted. It was changed to: "How severe is your back pain usually?" which is a combination of items 3 and 4 and hopefully easier to answer for elderly persons. The resulting questionnaire should be validated in patients with vertebral fractures and controls.

Before using GPCM, the polytomous Rasch model [also called partial credit model (PCM)] and the one-parameter logistic model (OPLM) were used to examine the fit of the domains [21, 22]. Both models were originally used in educational psychometrics, but their application in health care research is relatively new. The major drawback of the Rasch model is that it is often too restrictive to fit the data while it assumes identical discrimination parameters for all items [22, 23]. OPLM estimates item-difficulty parameters, as in the Rasch model, but imputes discrimination parameters for each item [22]. Although the overall R1C-statistic of the OPLM could be decreased for all domains in our study, indicating an increase of the fit, this did not influence the corresponding p value (data not shown). In education psychometrics, often, two answer categories are used, so it is possible that OPLM is less suitable for questionnaires with five or more categories, such as the Qualeffo. Therefore, we chose to analyze the data with GPCM. GPCM is the less restrictive version of PCM and was developed for polytomous data [12]. Using GPCM, not only the item difficulty parameters but also the item discrimination parameter is estimated by the program [23].

In conclusion, Qualeffo–31 was developed, which is a shorter version of Qualeffo–41, a questionnaire for assessing the quality of life in patients with vertebral deformities. Qualeffo–31 consists of three domains, with a reasonable to excellent fit to the GPCM. Although the fit to the GPCM supports the construct validity of the Qualeffo–31, validation in new data sets should be performed before using it in practice.

Acknowledgements This study is based on data from the Qualeffo validation study and the Multiple Outcomes of Raloxifene study. We would like to thank Susan Ewing for her advice. Natasja van Schoor was funded by a grant of Wyeth Research, Collegeville, Pennsylvania, USA.

Appendix

Pain

The four questions in this section regard the situation in the last week.

Question 1 (1)	Possible response
How often have you had back pain in the last week?	○ Never ○ 1 day per week or less ○ 2–3 days per week ○ 4–6 days per week ○ every day
Question 2 (2)	
If you have had back pain, for how long did you have back pain in the daytime?	○ Never ○ 1–2 hours ○ 3–5 hours ○ 6–10 hours ○ All day
Question 3 (3+4)	
How severe is your back pain usually?	○ No back pain ○ Mild ○ Moderate ○ Severe ○ Unbearable
Question 4 (5)	
Has the back pain disturbed your sleep in the last week?	○ Less than once per week ○ Once a week ○ Twice a week ○ Every other night ○ Every night

Physical function

The next 18 questions regard the present situation.

Question 1 (6)	Possible response
Do you have problems with dressing?	○ No difficulty ○ A little difficulty ○ Moderate difficulty ○ May need some help ○ Impossible without help
Question 2 (7)	
Do you have problems with taking a bath or shower?	○ No difficulty ○ A little difficulty ○ Moderate difficulty ○ May need some help ○ Impossible without help

Question 3 (8)	
Do you have problems with getting to or operating a toilet?	○ No difficulty ○ A little difficulty ○ Moderate difficulty ○ May need some help ○ Impossible without help

The next 5 questions also regard the present situation. If someone else does these things in your house, please answer as though you were responsible for them.

Question 4 (10)	
Can you do the cleaning?	○ Without difficulty ○ With a little difficulty ○ With moderate difficulty ○ With great difficulty ○ Impossible
Question 5 (11)	
Can you prepare meals?	○ Without difficulty ○ With a little difficulty ○ With moderate difficulty ○ With great difficulty ○ Impossible
Question 6 (12)	
Can you wash the dishes?	○ Without difficulty ○ With a little difficulty ○ With moderate difficulty ○ With great difficulty ○ Impossible
Question 7 (13)	
Can you do your day to day shopping?	○ Without difficulty ○ With a little difficulty ○ With moderate difficulty ○ With great difficulty ○ Impossible
Question 8 (14)	
Can you lift a heavy object of 20 lb (e.g., a crate of 12 bottles of milk, or a one-year-old child) and carry it for at least 10 yards?	○ Without difficulty ○ With a little difficulty ○ With moderate difficulty ○ With great difficulty ○ Impossible
Question 9 (15)	
Can you get up from a chair?	○ Without difficulty ○ With a little difficulty ○ With moderate difficulty ○ With great difficulty ○ Only with help
Question 10 (16)	
Can you bend down?	○ Easily ○ Fairly easily ○ Moderately ○ Very little ○ Impossible
Question 11 (17)	
Can you kneel down?	○ Easily ○ Fairly easily ○ Moderately ○ Very little ○ Impossible
Question 12 (18)	
Can you climb stairs to the next floor of a house?	○ Without difficulty ○ With a little difficulty ○ With at least one rest ○ With help only ○ Impossible
Question 13 (19)	
Can you walk 100 yards?	○ Fast without stopping ○ Slowly without stopping ○ Slowly with at least one stop ○ Only with help ○ Impossible
Question 14 (20)	
How often have you been outside in the last week?	○ Every day ○ 5–6 days/week ○ 3–4 days/week ○ 1–2 days/week ○ Less than once/week
Question 15 (21)	
Can you use public transport?	○ Without difficulty ○ With a little difficulty ○ With moderate difficulty ○ With great difficulty ○ Only with help
Question 16 (24)	
Can you do your gardening?	○ Yes ○ Yes with restrictions ○ Not at all ○ Not applicable
Question 17 (26)	
Can you visit a cinema, theatre, etc.?	○ Yes ○ Yes with restrictions ○ Not at all ○ No cinema, or theatre within a reasonable distance
Question 18 (27)	
How often did you visit friends or relatives during the last 3 months?	○ Once a week or more ○ Once or twice a month ○ Less than once a month ○ Never

Mental function

The next 9 questions regard the situation in the last week.

Question 1 (31)	Possible response
How would you rate your overall quality of life during the last week?	○ Excellent ○ Good ○ Satisfactory ○ Fair ○ Poor
Question 2 (33)	
Do you tend to feel tired?	○ In the morning ○ In the afternoon ○ Only in the evening ○ After strenuous activity ○ Almost never
Question 3 (34)	
Do you feel downhearted?	○ Almost every day ○ Three to five days a week ○ One or two days a week ○ Once in a while ○ Almost never

Question 4 (35)	
Do you feel lonely?	○ Almost every day ○ Three to five days a week ○ One or two days a week ○ Once in a while ○ Almost never
Question 5 (36)	
Do you feel full of energy?	○ Almost every day ○ Three to five days a week ○ One or two days a week ○ Once in a while ○ Almost never
Question 6 (37)	
Are you hopeful about your future?	○ Never ○ Rarely ○ Sometimes ○ Quite often ○ Always
Question 7 (38)	
Do you get upset over little things?	○ Never ○ Rarely ○ Sometimes ○ Quite often ○ Always
Question 8 (39)	
Do you find it easy to make contact with people?	○ Never ○ Rarely ○ Sometimes ○ Quite often ○ Always
Question 9 (40)	
Are you in good spirits most of the day?	○ Never ○ Rarely ○ Sometimes ○ Quite often ○ Always

References

1. Hasserius R, Redlund–Johnell I, Mellstrom D, Johansson C, Nilsson BE, Johnell O (2001) Vertebral deformation in urban Swedish men and women – Prevalence based on 797 subjects. Acta Orthop Scand 72:273–278
2. Jones G, White C, Nguyen T, Sambrook PN, Kelly PJ, Eisman JA (1996) Prevalent vertebral deformities: Relationship to bone mineral density and spinal osteophytosis in elderly men and women. Osteoporos Int 6:233–239
3. Lau EMC, Woo J, Chan H, Chan MKF, Griffith J, Chan YH et al (1998) The health consequences of vertebral deformity in elderly Chinese men and women. Calcif Tissue Int 63:1–4
4. Melton LJ III, Lane AW, Cooper C, Eastell R, O'Fallon WM, Riggs BL (1993) Prevalence and incidence of vertebral deformities. Osteoporos Int 3:113–119
5. Pluijm SM, Tromp AM, Smit JH, Deeg DJ, Lips P (2000) Consequences of vertebral deformities in older men and women. J Bone Miner Res 15:1564–1572
6. Silverman SL (1992). The clinical consequences of vertebral compression fracture. Bone 13 (Suppl 2):S27–31
7. Lips P, Cooper C, Agnusdei D, Caulin F, Egger P, Johnell O et al (1999) Quality of life in patients with vertebral fractures. Validation of the Quality of Life Questionnaire of the European Foundation for Osteoporosis (QUALEFFO). Osteoporos Int 10:150–160
8. Ross PD (1997) Clinical consequences of vertebral fractures. Am J Med 103:30S–42S
9. Lips P, Agnusdei D, Caulin F, Cooper C, Johnell O, Kanis J et al (1996) The development of a European questionnaire for quality of life in patients with vertebral osteoporosis. Scand J Rheumatol (Suppl) 103:84–85
10. Lips P, Cooper C, Agnusdei D, Caulin F, Egger P, Johnell O et al (1997) Quality of life as outcome in the treatment of osteoporosis: the development of a questionnaire for quality of life by the European Foundation for Osteoporosis. Osteoporos Int 7:36–38
11. Hays RD, Morales LS, Reise SP (2000) Item response theory and health outcomes measurement in the 21st century. Med Care 38:28–42
12. Cella D, Chang CH (2000) A discussion of item response theory and its applications in health status assessment. Med Care 38:66–72
13. Oleksik A, Lips P, Dawson A, Minshall ME, Shen W, Cooper C et al (2000) Health-related quality of life in postmenopausal women with low BMD with or without prevalent vertebral fractures. J Bone Miner Res 15:1384–1392
14. Jöreskog KG, Sörbom D (1996) Prelis 2 User's Reference Guide. Mooresville: SSI Scientific Software
15. Flora DB, Finkel EJ, Foshee VA (2003) Higher order factor structure of a self-control test: Evidence from confirmatory factor analysis with polychoric correlations. Educ Psychol Meas 63:112–127
16. Streiner DL (1994) Figuring out factors - the use and misuse of factor-analysis. Can J Psychiatry 39:135–140
17. Muraki E (1992) A generalized partial credit model-application of an EM algorithm. Applied Psychological Measurement 16:159–176
18. Glas CAW (1999) Modification indices for the 2–PL and the nominal response model. Psychometrika 64:273–294
19. Glas CAW, Falcon JCS (2003) A comparison of item-fit statistics for the three–parameter logistic model. Applied Psychological Measurement 27:87–106
20. van Buuren S, Hopman-Rock M (2001) Revision of the ICIDH Severity of Disabilities Scale by data linking and item response theory. Stat Med 20:1061–1076
21. Rasch G (1960) Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research
22. Verhelst ND, Glas CAW, Verstralen HHFM (1995) One-Parameter Logistic Model OPLM. Arnhem: CITO
23. Ware JE, Bjorner JB, Kosinski M (2000) Practical implications of item response theory and computerized adaptive testing–A brief summary of ongoing studies of widely used headache impact scales. Med Care 38:73–82