Measurement invariance of the WHOQOL-AGE questionnaire across three European countries

David Santos, Francisco J. Abad, Marta Miret, Somnath Chatterji, Beatriz Olaya, Katarzyna Zawisza, Seppo Koskinen, Matilde Leonardi, et al.

Quality of Life Research
An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation - Official Journal of the International Society of Quality of Life Research

ISSN 0962-9343
Volume 27
Number 4

Qual Life Res (2018) 27:1015-1025
DOI 10.1007/s11136-017-1737-8

Springer

Quality of Life Research (2018) 27:1015–1025
https://doi.org/10.1007/s11136-017-1737-8

Measurement invariance of the WHOQOL-AGE questionnaire across three European countries

David Santos[1] · Francisco J. Abad[1] · Marta Miret[2,3,4] · Somnath Chatterji[5] · Beatriz Olaya[3,6] · Katarzyna Zawisza[7] · Seppo Koskinen[8] · Matilde Leonardi[9] · Josep Maria Haro[3,6] · José Luis Ayuso-Mateos[2,3,4] · Francisco Félix Caballero[2,3,4]

Accepted: 8 November 2017 / Published online: 16 November 2017

Abstract

Purpose Developing valid and reliable instruments that can be used across countries is necessary. The present study aimed to test the comparability of quality of life scores across three European countries (Finland, Poland, and Spain).

Method Data from 9987 participants interviewed between 2011 and 2012 were employed, using nationally representative samples from the Collaborative Research on Ageing in Europe project. The WHOQOL-AGE questionnaire is a 13-item test and was employed to assess the quality of life in the three considered countries. First of all, two models (a bifactor model and a two-correlated factor model) were proposed and tested in each country by means of confirmatory factor models. Second, measurement invariance across the three countries was tested using multi-group confirmatory factor analysis for that model which showed the best fit. Finally, differences in latent mean scores across countries were analyzed.

Results The results indicated that the bifactor model showed more satisfactory goodness-of-fit indices than the two-correlated factor model and that the WHOQOL-AGE questionnaire is a partially scalar invariant instrument (only two items do not meet scalar invariance). Quality of life scores were higher in Finland (considered as the reference category: mean = 0, SD = 1) than in Spain (mean = − 0.547, SD = 1.22) and Poland (mean = − 0.927, SD = 1.26).

Conclusions Respondents from Finland, Poland, and Spain attribute the same meaning to the latent construct studied, and differences across countries can be due to actual differences in quality of life. According to the results, the comparability across the different considered samples is supported and the WHOQOL-AGE showed an adequate validity in terms of cross-country validation. Caution should be exercised with the two items which did not meet scalar invariance, as potential indicator of differential item functioning.

Keywords Quality of life · Measurement invariance · Multi-group confirmatory factor analysis · WHOQOL-AGE · Bifactor model

✉ Francisco Félix Caballero
felix.caballero@uam.es

[1] Department of Psychology, Universidad Autónoma de Madrid, Madrid, Spain

[2] Department of Psychiatry, Universidad Autónoma de Madrid, C/Arzobispo Morcillo 4, 28029 Madrid, Spain

[3] CIBER of Mental Health, Madrid, Spain

[4] Instituto de Investigación Sanitaria (IIS-Princesa), Hospital Universitario de La Princesa, Madrid, Spain

[5] Information, Evidence and Research, World Health Organization, Geneva, Switzerland

[6] Parc Sanitari Sant Joan de Déu, Universitat de Barcelona, Barcelona, Spain

[7] Department of Medical Sociology, Chair of Epidemiology and Preventive Medicine, Jagiellonian University Medical College, Krakow, Poland

[8] National Institute for Health and Welfare, Helsinki, Finland

[9] Fondazione IRCCS, Neurological Institute Carlo Besta, Milano, Italy

Introduction

The World Health Organization (WHO) argues that the measurement of health and the effects of health care must include not only an indication of changes in the frequency and severity of diseases, but also an estimation of the quality of life (QOL) related to health care [1]. QOL is a broad-ranging concept which can be thought of as "the individuals' perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns" [2].

The World Health Organization Quality Of Life (WHOQOL) Group [3] elaborated a series of instruments that seek to assess QOL, allowing for cross-cultural comparisons [4]. The rationale and assumption of the WHOQOL measurement instruments is that people themselves should be asked whether they are satisfied with their health and well-being: a person can be satisfied with his/her life, though disabled by disease or illness.

The first WHOQOL instrument [5], also called WHOQOL-100, was composed of 100 items and developed collaboratively over several years in a number of centers in diverse cultural settings. A brief version of the instrument, composed of 26 items and showing good psychometric properties, was also developed in order to reduce time and fatigue: the WHOQOL-BREF [6, 7]. Moreover, the WHOQOL-OLD [8] has been created as a specific WHOQOL module to assess quality of life in the elderly population, although needs to be administered together with WHOQOL-BREF. In order to reduce the application time, another WHOQOL instrument was created: the EUROHIS-QOL eight-item index [9], showing satisfactory internal consistency as well as good convergent and discriminant validity.

The WHOQOL-AGE is an instrument which has been designed specifically to assess QOL in aging populations and has been built upon previous WHOQOL instruments; it is based on the EUROHIS-QOL and the WHOQOL-OLD short form version 1 [10]. The WHOQOL-AGE has been developed within the Collaborative Research on Ageing in Europe (COURAGE in Europe) [11] project and it has been validated in samples from Finland, Poland, and Spain [12], using the same dataset that is considered in the present research and showing adequate psychometric properties. Since the application time of this instrument is shorter than the time necessary to complete other similar QOL scales specifically designed for the older population, it can be applied in large-scale population studies and busy clinical settings.

During the validation process of an instrument, it is not only necessary to test the reliability of the scores and other psychometric properties, but also to test whether the scores of the instrument are measurement invariant across groups (e.g., differences across countries, age groups, gender, etc.). According to Meredith and Millsap [13, 14], a measurement instrument is measurement invariant if an individual's probability of an observed score does not depend on his/her membership to a group, conditional on the true score. Measurement invariance in this sense assumes that the parameters of the measurement model which describes the relation between the latent variable and the observed scores are the same within the different groups.

Objectives of the current study

The present study aimed to test the comparability of the WHOQOL-AGE scores across different countries. First, the measurement invariance of the WHOQOL-AGE was assessed across three European samples (Finland, Poland, and Spain). And second, the latent scores in QOL were compared across these three countries, after assessing measurement invariance.

Method

Sample and design

Data were collected between 2011 and 2012, within the COURAGE in Europe project, an observational cross-sectional study comprising samples from the adult non-institutionalized population of Finland, Poland, and Spain. Specific details about the COURAGE in Europe project and the sampling strategy developed in the three countries are described elsewhere [11, 15].

The COURAGE in Europe project collected data on adults aged 50 years and older, plus a sample of adults aged between 18 and 49 years for comparison purposes. Among the participants older than 50, people older than 80 years were overrepresented in the sampling in order to avoid having small sample sizes for the oldest age groups and to allow for potential comparisons in subsequent follow-up studies. Before collecting data, informed consent was obtained from each participant. The research was approved by the local ethics research review boards [15].

Instrument

The WHOQOL-AGE comprises 13 items assessed on a five-point rating scale (items are shown in Appendix Table 7). The items which were not available in Finnish, Polish, or Spanish in previous WHOQOL instruments were translated from English into Finnish, Polish, and Spanish, following the World Health Organization's translation guidelines for

assessment instruments: a forward translation, a targeted back-translation, a review by a bilingual expert group, and a detailed translation report [16].

The response format is a combination of bipolar and unipolar formats. The unipolar format refers to scales that range from the absence of the attribute to its presence (e.g., response options from not at all to very much) while the bipolar format refers to scales using both the positively and the negatively worded response in the same item (e.g., response options from very negative to very positive) [17, 18]. The bipolar format was applied for items Q1 (ranging from 1 = "very bad" to 5 = "very good"), and items Q2–Q8 (ranging from 1 = "very dissatisfied" to 5 = "very satisfied"). The unipolar format was applied for items Q9–Q12 (ranging from 1 = "not at all" to 5 = "completely"), and Q13 (ranging from 1 = "not at all" to 5 = "an extreme amount").

In Caballero et al. [12], a second-order factorial structure was found for the 13 WHOQOL-AGE items, with the general factor representing QOL. Two first-order factors were also identified. Similar Cronbach's alpha values (0.88 and 0.84) were obtained for both factors in the overall sample, indicating a good internal consistency in the previous abovementioned study. Scoring information about the second-order general factor (QOL) and the two first-order factors found was also shown in Caballero et al. [12].

Statistical analysis

Socio-demographics

The socio-demographic characteristics of each sample (e.g., age, sex, years of education) were described, and differences across countries were analyzed using Chi-square test and analysis of variance (ANOVA). When large samples are considered, differences found can be due to the large sample sizes. Therefore, Cramer's *V* and partial eta-squared were computed to determine the effect size, assessing the magnitude according to Cohen's guidelines [19].

Effect indicators and causal indicators

QOL questionnaires often contain two different types of items [20–22]: effect indicators and causal indicators. A causal indicator is an empirical phenomenon that has an impact on the latent variable to be measured, whereas an effect indicator is an empirical phenomenon which is caused by the latent variable to be measured. For instance, in the case of QOL, the presence of a symptom of a disease (e.g., pain) should be better seen as a causal indicator (i.e., the presence of pain reduces QOL), whereas experiencing depression might be better conceived as an effect of low levels of QOL. When causal indicators are present, extra caution should be exerted before they are aggregated into a summated scale since they have a less uniform relationship with QOL. Increasing QOL is likely to affect all effect indicators, such as anxiety and depression, whereas will not change a causal indicator, such as the presence of pain.

Based on Fayers' approach [20], the global question about QOL in the WHOQOL-AGE (Q1—"How would you rate your quality of life?") was considered as a gold standard for the assessment of the latent variable QOL and the remaining 12 items were compared against Q1. For each of the 12 items, a contingency table between the item (e.g., Q2) and the external criterion considered (i.e., Q1) was obtained. For items that are causal indicators, an asymmetric relationship is expected: it is more probable to find people who score high in the indicator but low in quality of life than finding people who score low in the indicator but high in quality of life. A ratio (hereafter, the Indicator Identification Ratio, IIR) can be then computed. First, we obtain the number of people who score high (e.g., satisfied, very satisfied) in the indicator but low (e.g., bad or very bad) in quality of life (O_X) and the number of people who score low in the indicator but high in quality of life (O_Y). Second, the expected values of these frequencies (E_X and E_Y) were obtained, according to the independence null hypothesis. Finally, the IIR is obtained as the ratio of the two squared standardized residuals $\left(Z_X^2 / Z_Y^2\right)$. According to Fayers et al. [20], lower values in the IIR might indicate that the indicator is a cause rather than an effect indicator. Inter-item and adjusted item-test correlations were also calculated.

Confirmatory factor analysis

Two models were considered: (1) a bifactor model [23, 24], with one general factor influencing the items and two response format-specific factors (i.e., bipolar and unipolar), and (2) a two-correlated factor model. Both were tested in the overall sample, and separately by country (see Fig. 1). Factor loadings on specific factors were constrained to be positive in the bifactor model. The maximum likelihood mean-adjusted (MLM) estimator, based on maximum likelihood parameter estimates with standard errors and a mean-adjusted Chi-square test statistic robust to non-normality [25] was employed in each model.

Comparative Fit Index (CFI), Tucker–Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA) were considered to assess the goodness-of-fit of the models according to the cut-off points established in the literature [26, 27]: CFI > 0.90, TLI > 0.90, and RMSEA < 0.08. Akaike Information Criterion (AIC) [28] values were also reported.

Multi-group confirmatory factor analysis

Measurement invariance across the three countries was assessed by means of a Multi-group Confirmatory Factor

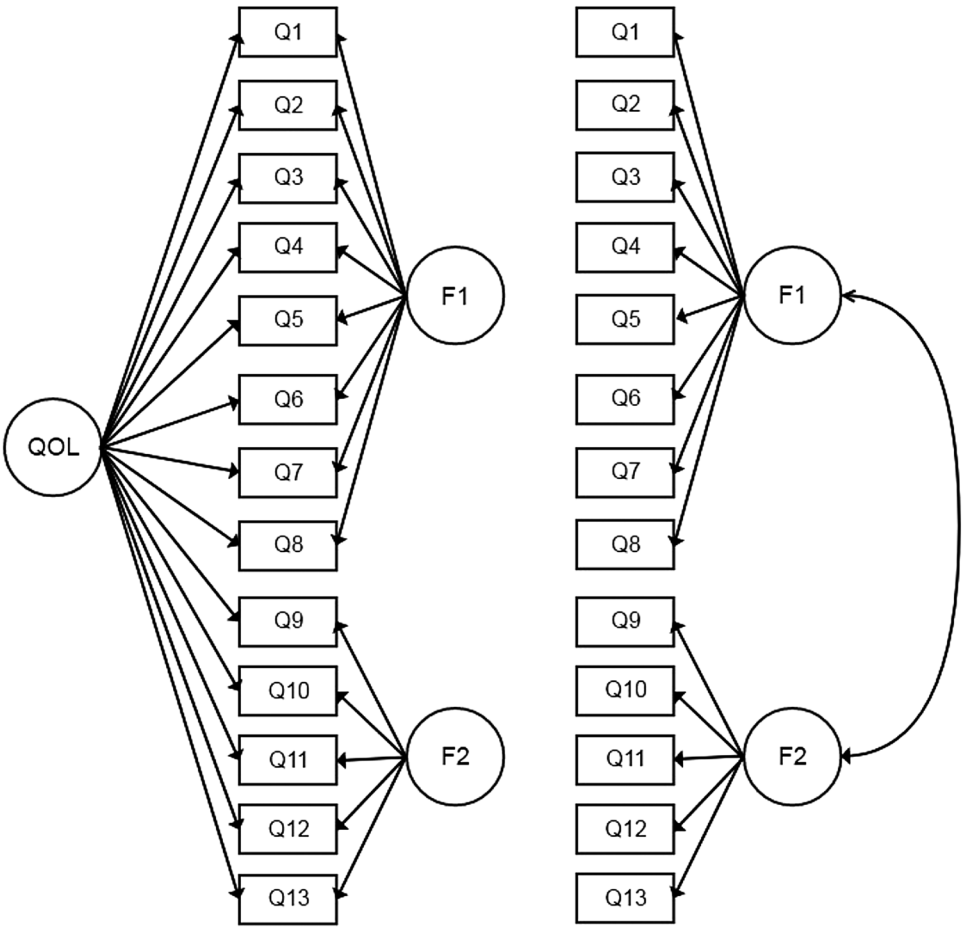

Fig. 1 Graphic representation of the bifactor model structure (left side) and the two-correlated factor model (right side)

Analysis (MG-CFA) for that model which showed a better fit to the data in the overall sample. A sequential constraint approach was employed to assess whether the same construct was measured across the Finnish, Polish, and Spanish samples. The invariance of the factor structure was studied comparing a set of nested models [29, 30]: (a) configural invariance model (factor structure is equal across groups, whereas the factor loadings, intercepts, and residual variances are allowed to differ across groups), (b) metric invariance model (the factor loadings are equal across groups), (c) scalar invariance model (the loadings and intercepts are constrained to be equal across groups), and (d) strict invariance model (the residual variances are also fixed to be equal across groups).

When configural invariance is reached, this suggests that the same items measure the constructs across countries. Metric invariance suggests that items share equivalent meaning across groups, in terms of the relationship with the factor [31], while scalar invariance suggests that differences in item means are due to differences in latent factor means across groups. Finally, when strict invariance is attained, the error variance associated to the items is the same across groups, suggesting that differences in means or covariances of the indicators are due to differences in latent factor distributions across groups [32].

The four invariance models were compared (i.e., configural vs. metric, metric vs. scalar, and scalar vs. strict) based on CFI values. The more restrictive model was considered as valid when $\Delta CFI < 0.01$ [33]. When the more restrictive model did not hold, then equality constraints which were specially imposed by the more restricted model were removed for one or more items until a $\Delta CFI < 0.01$ was achieved. Wald statistical tests were used to select these items. This procedure allowed us to detect the items that were responsible for the lack of invariance. The final model in which a subset of parameters is allowed to vary across groups is called a partial invariance model [34]. If partial invariance is tenable, groups can be compared at the latent level even if full measurement invariance is not attained.

Some constraints were added for identification purposes. Factor means and variances were fixed to zero and one, respectively, in Finland, considered as the reference group. When loadings were not constrained to be equal

across groups, latent factor variances were fixed to one for all groups; finally, when intercepts were not constrained to be equal across groups, latent factor means were fixed to zero in all groups.

Finally, the Wald test was used to assess differences in general factor latent scores on QOL across countries, when the level of scalar invariance was achieved. The mean and standard deviation (SD) in the general factor (QOL) were set to zero and one in Finland (reference group). Factor loadings and intercepts should be the same across countries for a correct interpretation of latent means across countries [30]. Higher latent scores indicated a better QOL. Cohen's *d* [19] associated to each pairwise comparison was computed. 95% confidence intervals for Cohen's *d* [35] were calculated using the MBESS package [36] in R. Mplus version 6 [37] was employed for structural equation modeling. The remaining analyses were computed using Stata 11 [38].

Results

Descriptive statistics

Mean age of the sample ($N = 9987$) was 58.10 years (SD = 16.70), with a 56.73% of women. The age varied between 18 and 100 years. The socio-demographics separated by country are shown in Table 1. Although significant differences were found across countries, the associated effect sizes were small (or moderate in the case of residential setting). Significant differences in the percentage of participants in each age group (18–49, 50–79, 80 + years) were found across countries ($p < 0.001$), with a higher rate of people aged between 18 and 49 years in Poland; however, this difference had associated a low effect size (Cramer's $V = 0.07$).

Effect indicators and causal indicators

Mean inter-item correlation for the WHOQOL-AGE items was 0.45, with pairwise correlations ranging from 0.27 to 0.76. As illustrated in Table 2, the adjusted item-test correlations ranged from 0.49 for Q7 to 0.74 for Q5. The items with the lower IIR are Q12 and Q13. Skewness, floor, and ceiling effects corresponding to the 13 WHOQOL-AGE items are also displayed (see Table 3).

Confirmatory factor analysis

For the bifactor model, satisfactory goodness-of-fit indices were obtained in Finland (RMSEA = 0.068; CFI = 0.931; TLI = 0.889), Poland (RMSEA = 0.056; CFI = 0.970; TLI = 0.951), and Spain (RMSEA = 0.057; CFI = 0.967; TLI = 0.946). The bifactor model showed a better fit than the two-correlated factor model in all cases (Table 4). The model fit in Finland was worse than in the other countries. In the total sample, the correlation between the factors of

Table 1 Socio-demographic characteristics of the sample ($N = 9987$) in each country

	Finland	Poland	Spain	*p* value	Effect size
Number of participants (*n*)	1845	3940	4202		
Sex, *n* (%)					
Female	1042 (56.48)	2370 (60.15)	2254 (53.64)	< 0.001	0.06
Male	803 (43.52)	1570 (39.85)	1948 (46.36)		
Current marital status, *n* (%)					
Not married	670 (36.31)	1739 (44.14)	1555 (37.01)	< 0.001	0.07
Married or in partnership	1175 (63.69)	2201 (55.86)	2647 (62.99)		
Residential setting, *n* (%)					
Rural	402 (21.79)	1702 (43.20)	564 (13.42)	< 0.001	0.31
Urban	1443 (78.21)	2238 (56.80)	3638 (86.58)		
Age, mean (SD)	58.13 (15.88)	57.02 (17.93)	59.10 (15.76)	< 0.001	0.01
Age group, *n* (%)					
18–49 years	477 (25.85)	1030 (26.14)	915 (21.78)	< 0.001	0.07
50–79 years	1187 (64.34)	2438 (61.88)	2977 (70.85)		
80 + years	181 (9.81)	472 (11.98)	310 (7.38)		
Years of education, mean (SD)	12.38 (4.21)	11.65 (3.99)	10.91 (6.32)	< 0.001	0.01

All differences were found significant at the 99% confidence level. Effect size: Cramer's *V* for χ^2 test (categorical variables) and partial eta-squared (η^2) for ANOVA test (quantitative variables). Effect size was reported for all the differences that were found to be significant at the 95% confidence level. Cramer's *V* values of 0.10, 0.30, and 0.50, constitute small, medium, and large effect sizes. whereas these values are 0.01, 0.06, and 0.14, respectively, for partial eta-squared

Table 2 Procedures to distinguish between effect indicators and causal indicators

Item	Proportion of people who score high in the indicator among people with low quality of life	Proportion of people who score low in the indicator among people with high quality of life	Indicator identification ratio	r_{jX1}	r_{jX}
Q2	0.53	0.04	0.43	0.36	0.58
Q3	0.27	0.06	0.55	0.46	0.70
Q4	0.43	0.02	0.66	0.42	0.71
Q5	0.35	0.03	0.54	0.47	0.74
Q6	0.57	0.01	0.58	0.38	0.66
Q7	0.64	0.02	0.43	0.30	0.49
Q8	0.41	0.04	0.84	0.36	0.61
Q9	0.33	0.03	0.52	0.44	0.73
Q10	0.36	0.03	0.53	0.40	0.70
Q11	0.28	0.04	0.49	0.43	0.71
Q12	0.28	0.12	0.27	0.38	0.50
Q13	0.48	0.04	0.39	0.34	0.53

r_{jX1} item-Q1 correlation, r_{jX} adjusted item-test correlation

Table 3 Skewness, floor, and ceiling effects for the items of the WHOQOL-AGE questionnaire

Item	Skewness	Floor	Ceiling
Q1	−0.68	0.01	0.12
Q2	−1.01	0.01	0.21
Q3	−0.83	0.02	0.17
Q4	−0.99	0.01	0.20
Q5	−1.07	0.02	0.23
Q6	−0.99	0.01	0.22
Q7	−1.13	0.01	0.25
Q8	−0.93	0.01	0.15
Q9	−0.78	0.02	0.31
Q10	−0.74	0.01	0.28
Q11	−0.66	0.03	0.19
Q12	−0.43	0.07	0.19
Q13	−0.88	0.03	0.29

the two-correlated factor model was 0.810 [95% CI (0.786, 0.826)].

Multi-group confirmatory factor analysis

As shown in Table 5, the MG-CFA displayed an adequate fit for the configural invariance model: RMSEA = 0.058; CFI = 0.964; TLI = 0.941. The metric invariance model also showed an adequate fit according to goodness-of-fit indices (RMSEA = 0.056; CFI = 0.957; TLI = 0.947). The difference in CFI values between configural and metric invariance models was < 0.01 (ΔCFI = 0.007). Note that constraining the loadings for the specific factors did not decrease the model fit (ΔCFI = 0.002). More importantly, invariance of loadings on the QOL general factor was also tenable (ΔCFI = 0.005). Regarding the scalar invariance model, an adequate fit was found (RMSEA = 0.062; CFI = 0.941; TLI = 0.935), but the change in the CFI value was higher than 0.01 (ΔCFI = 0.016) when comparing metric and scalar invariance models. As the full invariance model was not reached, we tested partial invariance models in order to detect the items that were responsible for the lack of invariance. Based on the Wald test, we sequentially dropped the equality constraint of the intercept for items Q12 (Model 3.1.) and Q13 (Model 3.2.), until the criterion based on a change in the CFI value (ΔCFI < 0.01) was achieved. Finally, the strict invariance model displayed satisfactory fit according to the goodness-of-fit indices considered (RMSEA = 0.059; CFI = 0.941; TLI = 0.941) and produced a small change in the CFI value (ΔCFI = 0.007) when compared with the partial scalar model. Thus, strict invariance across countries was achieved after unconstraining the intercepts of items Q12 and Q13.

The standardized loadings for the configural model and the strict invariance model are shown in Table 6 for the three countries. Regarding the configural model, the standardized loadings of the 13 WHOQOL-AGE items on the general factor ranged from 0.391 to 0.738 in Finland, from 0.429 to 0.824 in Poland, and from 0.314 to 0.831 in Spain. Similar ranges for each country were found when running the strict invariance model (Table 6).

Finally, latent mean scores on QOL across countries were calculated for the general factor based on the partial scalar invariance model. Taking Finland as reference group (mean = 0, SD = 1), significantly lower latent scores in QOL were found in Spain (mean = −0.547, SD = 1.22) and Poland (mean = −0.927, SD = 1.26). The difference in QOL latent scores between Finland and Poland had a large effect size associated [Cohen's d = 0.78; 95% CI = (0.72, 0.84)]. A moderate effect size was associated with the differences between

Table 4 Goodness-of-fit indices associated to the bifactor and the two-correlated factor models

Goodness-of-fit	Total sample	Finland	Poland	Spain
Bifactor model				
χ^2 (d.f.)	1431.13 (48)	454.90 (48)	638.35 (48)	692.98 (48)
RMSEA (90% CI)	0.054 (0.051, 0.056)	0.068 (0.062, 0.074)	0.056 (0.052, 0.060)	0.057 (0.053, 0.060)
AIC	266646.879	48337.121	104041.802	110018.620
CFI	0.971	0.931	0.970	0.967
TLI	0.952	0.889	0.951	0.946
Two-correlated factor model				
χ^2 (d.f.)	3059.23 (60)	614.63 (60)	1252.28(60)	1503.74 (60)
RMSEA (90% CI)	0.071 (0.069, 0.073)	0.071 (0.066, 0.076)	0.071 (0.068, 0.074)	0.076 (0.072, 0.079)
AIC	268868.742	48524.024	104882.777	111142.745
CFI	0.936	0.907	0.940	0.925
TLI	0.917	0.879	0.921	0.903
Correlation between factors (95% CI)	0.810 (0.786, 0.826)	0.880 (0.827, 0.933)	0.791 (0.764, 0.818)	0.759 (0.726, 792)

All Chi-square values were significant at the 99.9% confidence level

Table 5 Goodness-of-fit indices for the different invariance bifactor models across countries

	Number of parameters	RMSEA (90% CI)	AIC	CFI	ΔCFI	TLI
Model 1: configural	168	0.058 (0.056, 0.061)	262397.359	0.964	–	0.941
Model 2a: metric (specific factors)	146	0.056 (0.054, 0.058)	262445.671	0.962	0.002	0.946
Model 2b: metric (general factor)	122	0.056 (0.053, 0.058)	262761.248	0.957	0.005	0.947
Model 3: scalar	102	0.062 (0.060, 0.064)	263707.770	0.941	0.016*	0.935
Model 3.1: partial scalar [μ_{12}]	104	0.061 (0.059, 0.063)	263566.392	0.944	0.013*	0.937
Model 3.2: partial scalar [μ_{12}, μ_{13}]	106	0.059 (0.057, 0.061)	263301.548	0.948	0.009	0.941
Model 4: strict	80	0.059 (0.057, 0.061)	263902.556	0.941	0.007	0.941

Parameters which were considered as free to assess partial scalar invariance are given in square brackets

μ_{12} intercept of Q12, μ_{13} intercept of Q13

*ΔCFI ≥ 0.01 regarding the previous invariant model (configural/metric/partial scalar)

Spain and Poland [Cohen's $d = 0.34$; 95% CI = (0.30, 0.38)], and between Finland and Spain [Cohen's $d = 0.51$; 95% CI = (0.45, 0.57)].

Discussion

The present study was focused on a questionnaire specifically designed for assessing QOL in aging population, the WHOQOL-AGE [12]. The main aim of the current research was to test the comparability of WHOQOL-AGE scores across three European countries. The WHOQOL-AGE was found to be partially invariant across the Finland, Poland, and Spain.

A bifactor model is proposed in the present study. The two specific factors comprised, respectively, those items with bipolar and unipolar item response formats. The 13 WHOQOL-AGE items are loaded on the QOL general factor. The items assigned to each specific factor slightly differ from those identified in Caballero et al. [12]. In that study, the general item Q1 was considered as belonging to both specific factors, while in the present study, this first item was assigned only to the first one, based on the response format and the structure of the bifactor model.

This research provides evidence for the partial measurement invariance of the WHOQOL-AGE across Finland, Poland, and Spain. Regarding the measurement invariance, two item intercepts (Q12 and Q13) were unconstrained in

Table 6 Standardized loadings for the configural model and the strict invariance model, in the three countries

	Configural general			Configural specific			Strict general			Strict specific		
	F	P	S	F	P	S	F	P	S	F	P	S
Q1	0.632	0.525	0.580	0.000	0.080	0.044	0.508	0.594	0.579	0.000	0.000	0.000
Q2	0.513	0.706	0.661	0.021	0.024	0.000	0.588	0.673	0.658	0.000	0.000	0.000
Q3	0.703	0.811	0.819	0.000	0.000	0.000	0.722	0.794	0.782	0.000	0.000	0.000
Q4	0.672	0.731	0.722	0.342	0.427	0.449	0.680	0.751	0.732	0.408	0.383	0.409
Q5	0.738	0.824	0.831	0.047	0.117	0.041	0.776	0.839	0.829	0.007	0.006	0.007
Q6	0.483	0.649	0.624	0.352	0.541	0.577	0.599	0.676	0.655	0.444	0.426	0.452
Q7	0.396	0.491	0.427	0.312	0.356	0.400	0.396	0.471	0.453	0.373	0.377	0.398
Q8	0.548	0.569	0.629	0.452	0.440	0.369	0.534	0.615	0.594	0.406	0.397	0.420
Q9	0.690	0.698	0.671	0.141	0.296	0.331	0.634	0.696	0.672	0.196	0.293	0.338
Q10	0.529	0.634	0.638	0.370	0.332	0.342	0.573	0.639	0.614	0.202	0.306	0.353
Q11	0.583	0.658	0.591	0.387	0.484	0.519	0.604	0.641	0.606	0.311	0.449	0.508
Q12	0.395	0.458	0.314	0.343	0.416	0.540	0.362	0.407	0.381	0.297	0.454	0.509
Q13	0.391	0.429	0.466	0.177	0.267	0.403	0.403	0.464	0.441	0.219	0.343	0.391

The differences in factor loadings across countries for the strict invariance model are due to the variability in the latent factors. According to the bifactor structure, item loadings on the general factor and in both specific factors are shown for the two models

F Finland, *P* Poland, *S* Spain

order to reach partial scalar invariance. After freeing these intercepts, the strict measurement invariance model showed adequate goodness-of-fit. For the two items which showed potential differential item functioning (DIF; Q12 and Q13), higher intercepts were found in Finland, followed by Spain and Poland. Different socio-economic characteristics in each country can be influencing the different performance of Q12 across countries. In the case of Q13, a question asking for satisfaction with intimate relationships, the presence of missing values, which could not be missing at random, can be related to potential DIF found across countries. Although the percentage of missing values in Q13 was not too high in any of the three countries, there were more people in Spain (8.3%) that did not respond to this item, as described in the previous study by Caballero et al. [12]. Further research could explore potential cultural differences that might make this a sensitive question. Since Q13 adds valuable information different to other questions, the possibility of dropping it was not considered in the present study.

A potential problem which can be found when dealing with QOL questionnaires is the presence of causal indicators instead of effect indicators. Causal indicators may cause reduction in QOL for those subjects experiencing them, but the reverse relationship need not apply. Since factor analysis is based upon analyzing the correlation matrix and assuming all items to be effect indicators, procedures based on factor analysis could be largely irrelevant as a method of scale validation for those QOL instruments that contain causal indicators [22]. Another consequence of causal relationships that we can exploit is that inter-item correlations do not reflect the latent variable. Thus, in many situations, causal variables will give rise to seemingly unexplainable factor structures [21].

The method developed by Fayers et al. [20] has been employed in the present manuscript in order to identify potential causal indicators. Although the results of the procedure and the moderate to strong item-test correlation coefficients could suggest that the WHOQOL-AGE items are effect indicators, the findings should be interpreted with caution, since there is no statistical test associated to the method and the judgment becomes a very subjective one in interpreting the pattern of responding [39]. Additionally, although low correlations of an indicator with the remaining items could suggest that an item might be a causal indicator, we cannot discard that items with a high correlation with the remaining ones would still be a causal indicator. It is also hard to establish a direction of causality based on cross-sectional data. For example, one could argue that having high QOL is not causing you to have enough money to meet your needs. Although we agree that this is arguable, one could also think that having high QOL does not cause you to have more money but it can cause you to perceive the money in a different way.

One of the main strengths of the present study is that samples from Finland, Poland, and Spain were used, representing several geographic regions in Europe (Northern, Eastern, and Southern Europe) [40] and different social welfare systems [41]. A second strength is that the measurement

invariance tested in the present study provided evidence for using the WHOQOL-AGE to compare QOL scores across countries. Moreover, the proposed bifactor model has shown satisfactory goodness-of-fit indices in the three considered countries, supporting the idea of a general factor of QOL with two underlying specific factors.

One of the results of our analysis is that measurement invariance was not met without freeing some parameters to be estimated independently among the three samples. After two items were unconstrained (Q12 and Q13), partial scalar invariance was met and the next model could be tested. Moreover, it was common in previous applied research testing multi-group invariance to free some parameters in order to reach the measurement invariance models [42, 43]. It is also important to note that configural and metric invariance were met without freeing any parameters. Another potential challenge of the present research deals with using methods for continuous data with categorical items. Methods for continuous data (such as MLR) often display greater power to detect scalar non-invariance, but lower power to identify metric non-invariance when compared to methods for categorical data (such as WLSMV) [44]. Additionally, previous authors have shown that categorical estimators have two limitations when applied to testing invariance: (a) the use of "latent response variates" as intermediate variables operating between the ordered categorical indicators and the factors raise some complexities [45]; (b) for categorical estimators, robust population-corrected statistics (i.e., CFI) usually computed by traditional SEM software (e.g., Mplus) might be problematic since these are highly dependent on the distribution of thresholds [46].

Finally, some lines of future research can be proposed related to this study. Since the WHOQOL-AGE is an instrument specifically designed for aging population, other measurement invariance models can be proposed to assess whether there are differences in scores between the old and the oldest old populations. Moreover, differences based on gender could be explored. Although the WHOQOL-AGE is a recently developed instrument, some studies have just used this instrument to assess determinants of QOL [47–49], while other studies [50, 51] have mentioned the psychometric characteristics found in Caballero et al. [12]. The present study is the first measurement invariance analysis conducted for the WHOQOL-AGE.

To summarize, the WHOQOL-AGE questionnaire has shown a partial measurement invariance across Finland, Poland, and Spain. The instrument can be used in these countries to assess QOL because the difference between scores in the three countries might be attributable to actual differences in quality of life rather than other characteristics of the scale (e.g., item comprehension or familiarity with item response formats).

Appendix

See Table 7.

Table 7 The WHOQOL-AGE questionnaire

Q1. How would you rate your quality of life?
Q2. How satisfied are you with your hearing, vision, or other senses overall?
Q3. How satisfied are you with your health?
Q4. How satisfied are you with yourself?
Q5. How satisfied are you with your ability to perform your daily living activities?
Q6. How satisfied are you with your personal relationships?
Q7. How satisfied are you with the conditions of your living place (your home)?
Q8. How satisfied are you with the way you use your time?
Q9. Do you have enough energy for everyday life?
Q10. How much control do you have over the things you like to do?
Q11. To what extent are you satisfied with your opportunities to continue achieving in life?
Q12. Do you have enough money to meet your needs?
Q13. How satisfied are you with your intimate relationships in your life?

All the response options use a five-point rating scale, ranging from very bad to very good for Q1, from very dissatisfied to very satisfied for Q2–Q8, from not at all to completely for Q9–Q12, and from not at all to an extreme amount for Q13

Acknowledgements The present research has been funded by the Seventh Framework Programme of the European Commission (FP7/2007-2013) under Grant Agreement Number 223071 (COURAGE in Europe), by the Instituto de Salud Carlos III-FIS research Grant Numbers PS09/00295 and PS09/01845, by the Spanish Ministry of Science and Innovation's ACI-Promociona (ACI2009-1010), and the Mental Health and Disability Instrument Library Platform (CIBERSAM). The study was also supported by the Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Instituto de Salud Carlos III. D.S. is grateful to the Universidad Autónoma de Madrid for the doctoral fellowship (Reference No. FPI-UAM2015). F.J.A. is grateful to the Ministerio de Economia y Competitividad (Grant PSI2013-44300-P). All authors gratefully acknowledge the input of Prof. Mick Power during the process of selecting the WHOQOL-AGE items.

References

1. World Health Organization. (1997). *WHOQOL: Measuring quality of life*. Geneva: World Health Organization.
2. The WHOQOL Group. (1995). The World Health Organization quality of life assessment (WHOQOL): Position paper from the World Health Organization. *Social Science & Medicine, 41*(10), 1403–1409.
3. The WHOQOL Group. (1996). What quality of life? World Health Organization quality of life assessment. *World Health Forum, 17*(4), 354–356.
4. Skevington, S. M. (2002). Advancing cross-cultural research on quality of life: observations drawn from the WHOQOL development. *Quality of Life Research, 11*(2), 135–144.
5. The WHOQOL Group. (1998). The World Health Organization quality of life assessment (WHOQOL): Development and general psychometric properties. *Social Science & Medicine, 46*(12), 1569–1585.
6. The WHOQOL Group (1998). Development of the World Health Organization WHOQOL-BREF quality of life assessment. *Psychological Medicine, 28*(03), 551–558.
7. Skevington, S. M., Lotfy, M., & O'Connell, K. A. (2004). The World Health Organization's WHOQOL-BREF quality of life assessment: Psychometric properties and results of the international field trial. A report from the WHOQOL group. *Quality of Life Research, 13*(2), 299–310.
8. Power, M., Quinn, K., & Schmidt, S. (2005). Development of the WHOQOL-old module. *Quality of Life Research, 14*(10), 2197–2214.
9. Schmidt, S., Mühlan, H., & Power, M. (2006). The EUROHIS-QOL 8-item index: psychometric results of a cross-cultural field study. *The European Journal of Public Health, 16*(4), 420–428.
10. Fang, J., Power, M., Lin, Y., Zhang, J., Hao, Y., & Chatterji, S. (2012). Development of short versions for the WHOQOL-OLD module. *The Gerontologist, 52*(1), 66–78.
11. Leonardi, M., Chatterji, S., Koskinen, S., Ayuso-Mateos, J. L., Haro, J. M., Frisoni, G., et al. (2014). Determinants of health and disability in ageing population:the COURAGE in Europe project (Collaborative Research on Ageing in Europe). *Clinical Psychology & Psychotherapy, 21*(3), 193–198.
12. Caballero, F. F., Miret, M., Power, M., Chatterji, S., Tobiasz-Adamczyk, B., Koskinen, S., et al. (2013). Validation of an instrument to evaluate quality of life in the aging population: WHOQOL-AGE. *Health and Quality of Life Outcomes, 11*, 177.
13. Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Pyschometrika, 58*, 525–543.
14. Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57*(2), 289–311.
15. Miret, M., Caballero, F. F., Chatterji, S., Olaya, B., Tobiasz-Adamczyk, B., Koskinen, S., et al. (2014). Health and happiness: cross-sectional household surveys in Finland, Poland and Spain. *Bulletin of the World Health Organization, 92*(10), 716–725.
16. World Health Organization. (2013). *Process of translation and adaptation of instruments*. Geneva: World Health Organization. Retrieved from http://www.who.int/substance_abuse/research_tools/translation/en/.
17. Rey, J. J., Abad, F. J., Barrada, J. R., Garrido, L. E., & Ponsoda, V. (2014). The impact of ambiguous response categories on the factor structure of the GHQ–12. *Psychological Assessment, 26*(3), 1021–1030.
18. Schweizer, K., & Schreiner, M. (2010). Avoiding the effect of item wording by means of bipolar instead of unipolar items: An application to social optimism. *European Journal of Personality, 24*(2), 137–150.
19. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.
20. Fayers, P. M., Hand, D. J., Bjordal, K., & Groenvold, M. (1997). Causal indicators in quality of life research. *Quality of Life Research, 6*(5), 393–406.
21. Fayers, P. M., & Hand, D. J. (2002). Causal variables, indicator variables and measurement scales: an example from quality of life. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 165*(2), 233–253.
22. Fayers, P. M., & Hand, D. J. (1997). Factor analysis, causal indicators and quality of life. *Quality of Life Research, 6*, 139–150.
23. Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*, 544–559.
24. Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*(1), 19–31.
25. Hox, J. J., Mass, C. J. M., & Brinkhuis, J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica, 64*, 157–170.
26. Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.
27. Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552–566.
28. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions On Automatic Control, 19*, 716–723.
29. Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *MedicalCare, 44*(11 Suppl 3), S78–S94.
30. van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486–492.
31. Emerson, S. D., Guhn, M., & Gadermann, A. M. (2017). Measurement invariance of the Satisfaction with Life Scale: Reviewing three decades of research. *Quality of Life Research, 26*(9) 1–14.
32. Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling, 14*(3), 435–463.
33. Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255.

34. Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 3*(105), 456–466.
35. Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*(4), 532–574.
36. Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software, 20*(8), 1–24.
37. Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide: Statistical analysis with latent variables: User's guide*. Los Angeles: Muthén & Muthén.
38. StataCorp (2011). *Stata statistical software: Release 12*. College Station, TX: StataCorp LP.
39. Streiner, D. L. (2003). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment, 80*(3), 217–222.
40. United Nations (2013). *Composition of macro geographical (continental) regions, geographical sub-regions, and selected economic and other groupings*. Retrieved August 1, 2013, from http://unstats.un.org/unsd/methods/m49/m49regin.htm.
41. Eikemo, T. A., Huisman, M., Bambra, C., & Kunst, A. E. (2008). Health inequalities according to educational level in different welfare regimes: A comparison of 23 European countries. *Sociology of health & Illness, 30*(4), 565–582.
42. Abad, F. J., Sorrel, M. A., Román, F. J., & Colom, R. (2016). The relationships between WAIS-IV factor index scores and educational level: A bifactor model approach. *Psychological Assessment, 28*(8), 987–1000.
43. Mellor-Marsá, B., Miret, M., Abad, F. J., Chatterji, S., Olaya, B., Tobiasz-Adamczyk, B., et al. (2016). Measurement invariance of the day reconstruction method: Results from the COURAGE in Europe project. *Journal of Happiness Studies, 17*(5), 1769–1787.
44. Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(2), 167–180.
45. Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York: Routledge.
46. Xia, Y. (2016). Investigating the chi-square-based model-fit indexes for WLSMV and ULSMV estimators. Doctoral dissertation, The Florida State University.
47. Lara, E., Koyanagi, A., Caballero, F., Domènech-Abella, J., Miret, M., Olaya, B., et al. (2017). Cognitive reserve is associated with quality of life: A population-based study. *Experimental Gerontology, 87*, 67–73.
48. Raggi, A., Corso, B., Minicuci, N., Quintas, R., Sattin, D., De Torres, L., et al. (2016). Determinants of quality of life in ageing populations: Results from a cross-sectional study in Finland, Poland and Spain. *PLoS ONE, 11*(7), e0159293.
49. Garin, N., Olaya, B., Moneta, M. V., Miret, M., Lobo, A., Ayuso-Mateos, J. L., & Haro, J. M. (2014). Impact of multimorbidity on disability and quality of life in the Spanish older population. *PLoS ONE, 9*(11), e111498.
50. Snell, D. L., Siegert, R. J., Surgenor, L. J., Dunn, J. A., & Hooper, G. J. (2016). Evaluating quality of life outcomes following joint replacement: Psychometric evaluation of a short form of the WHOQOL-Bref. *Quality of Life Research, 25*(1), 51–61.
51. Torisson, G., Stavenow, L., Minthon, L., & Londos, E. (2016). Reliability, validity and clinical correlates of the Quality of Life in Alzheimer's disease (QoL-AD) scale in medical inpatients. *Health and Quality of Life Outcomes, 14*, 90.