ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

Themed Section: EQ-HWB

Developing a New Generic Health and Wellbeing Measure: Psychometric Survey Results for the EQ-HWB

Tessa Peasgood, PhD, Clara Mukuria, PhD, John Brazier, PhD, Ole Marten, MSc, Simone Kreimeier, PhD, Nan Luo, PhD, Brendan Mulhern, PhD, Wolfgang Greiner, PhD, A. Simon Pickard, PhD, Federico Augustovski, PhD, Lidia Engel, PhD, Luz Gibbons, MSc, Zhihao Yang, PhD, Andrea L. Monteiro, MSc, Maja Kuharic, MSc, Maria Belizan, MSc, Jakob Bjørner, PhD

ABSTRACT

Objectives: The development of measures such as the EQ-HWB (EQ Health and Wellbeing) requires selection of items. This study explored the psychometric performance of candidate items, testing their validity in patients, social carer users, and carers.

Methods: Article and online surveys that included candidate items (N = 64) were conducted in Argentina, Australia, China, Germany, United Kingdom, and the United States. Psychometric assessment on missing data, response distributions, and known group differences was undertaken. Dimensionality was explored using exploratory and confirmatory factor analysis. Poorly fitting items were identified using information functions, and the function of each response category was assessed using category characteristic curves from item response theory (IRT) models. Differential item functioning was tested across key subgroups.

Results: There were 4879 respondents (Argentina = 508, Australia = 514, China = 497, Germany = 502, United Kingdom = 1955, United States = 903). Where missing data were allowed, it was low (UK article survey 2.3%; US survey 0.6%). Most items had responses distributed across all levels. Most items could discriminate between groups with known health conditions with moderate to large effect sizes. Items were less able to discriminate across carers. Factor analysis found positive and negative measurement factors alongside the constructs of interest. For most of the countries apart from China, the confirmatory factor analysis model had good fit with some minor modifications. IRT indicated that most items had well-functioning response categories but there was some evidence of differential item functioning in many items.

Conclusions: Items performed well in classical psychometric testing and IRT. This large 6-country collaboration provided evidence to inform item selection for the EQ-HWB measure.

Keywords: EQ-HWB, health and wellbeing, item selection, item response theory, measurement development, psychometrics, quality-adjusted life-year.

VALUE HEALTH. 2022; 25(4):525–533

Introduction

The "Extending the QALY" project aimed to develop a new generic measure, the EQ-HWB (EQ Health and Wellbeing), that can be used in economic evaluation across health, social care, and public health to estimate quality-adjusted life-years, based on the views of users and beneficiaries of these services including informal carers. The aim was to develop a long and short version, with the latter designed to be amenable to valuation to address the need for a single measure that would be used within and across the different beneficiaries.[1]

Stage 1 of this project established the domains for the measure, which were based on aspects of health and wellbeing identified as important in qualitative research by future users of the measure.[2] A qualitative literature review was undertaken to identify qualitative reviews on the impact on health and wellbeing of health conditions, being an informal carer or being a social care user, and primary qualitative work used in measure development.[3] This resulted in 32 subdomains organized into 7 high level domains (activity, autonomy, cognition, feelings and emotions, relationships, physical sensations, and self-identity).

Stage 2 generated a list of candidate items for each subdomain and stage 3 explored the content and face validity of these items (n = 97) using a standardized interview protocol across 6 countries (Argentina, Australia, China, England, Germany, and the United States) with individuals with various physical and mental health conditions, carers, and social care users.[4] This explored potential users' interpretation and views of the items. During stage 2, criteria for item selection were developed that reflected the aims of the project, and these were taken into consideration at every stage to support item selection (these criteria are discussed in a separate article[5]). These criteria aimed to identify items which

would work well for both measurement and valuation (such as brief and unambiguous items). Findings from the interviews were used to identify items (n = 64) to take forward to the psychometric assessment.

Psychometric methods are widely used in the development of outcome measures and are an essential step in generating a valid and reliable questionnaire.[6] Current best practice recommends a combination of classical test theory approaches, confirmatory factor analysis (CFA), and analysis based on modern measurement theory such as item response theory (IRT) or Rasch analysis.[7] Stage 2 provided qualitative evidence on the face validity of the proposed items whereas the stage 4 assessment aimed to provide quantitative evidence of validity in a larger sample. This article sets out the methods, results, and discussion of stage 4: classical psychometric analyses, factor analyses, and IRT analyses.

Methods

Future users of the new measure include patients, social care users, and informal carers. Therefore, a survey was undertaken targeting these groups and healthy individuals in 6 countries (Argentina, Australia, China, Germany, the United Kingdom, and the United States). Informal carers were defined as those who looked after friends or family because they were sick, disabled, or elderly. The study was initiated in the United Kingdom, and then other researchers who were members of the EuroQol Group were invited to apply to replicate aspects of the study including face validity (stage 2) and this stage. Applications were assessed by the funder with focus on a mix of countries with consideration of different languages and potential cultural differences.

Sample

All participants were aged 18 years or older and able to complete a questionnaire in the main language of their country. In the United Kingdom, patients were recruited from online panels (target n = 1200) and from National Health Service (NHS) Trusts and primary care (target n = 800) with the latter invited by health or other professionals in person or by post with a self-complete article questionnaire. The invite included an option to request interviewer assistance (to encourage frail adults to take part) or to complete the survey online. For the online UK panel, patients with cancer, depression or anxiety, asthma or chronic obstructive pulmonary disease, diabetes, arthritis, heart conditions, or irritable bowel syndrome/Crohn's disease were targeted ($n \geq 100$ in the online panel). Conditions were selected to represent both long-term physical and mental health conditions including those with the highest prevalence.[8] Social care users and informal carers were recruited via both the NHS and online.

Participants in the other countries were recruited from online panels only, targeting different groups that were dependent on the context: Argentina (cancer, mental health, diabetes, carers), Australia (mental health, people experiencing pain, carers, people who use care aids), China (depression, generalized anxiety disorder, chronic hepatitis B, human immunodeficiency virus/acquired immunodeficiency syndrome, and carers), Germany (cancer, carers), and the United States (cancer). A healthy general population (defined as a visual analog score on health of >80) was also recruited online in all countries.

For both factor and IRT analysis, a sample size of 500 is considered adequate.[9] To ensure an adequate sample was achieved, the overall target sample size in the United Kingdom was 2000 participants across different clusters, age groups, and diagnoses to allow for subgroup analysis where necessary whereas for the other countries the target was 500 with fewer subgroups.

Data Collection

The selection of items and additional questionnaires for inclusion in the survey questionnaire balanced respondent burden with data requirements for the analysis. A survey was designed that included the EQ-HWB candidate items (n = 64). Domain and subdomain analysis to explore the performance of items is more robust if there are at least 4 items per subdomain[10] so 2 additional items were included to facilitate the IRT analysis: "I felt cheerful," which was taken from the Warwick-Edinburgh Mental Wellbeing Scale[11] and included in the happiness/sadness domain, and "How difficult is it for you to wash, toilet, dress yourself, eat or care for your appearance?" (with different response options to those tested in face validity), which was taken from the AQol-8D[12] and included in the personal care domain. Positive and negative items were retained from stage 3 despite the acknowledgment that combining both positively and negatively worded items would be challenging for preference elicitation exercises.

Response options for most of the items used frequency terms (not at all, only occasionally, some of the time, often, most, or all of the time [n = 55]) with some items using severity terms (mild, slight, moderate, severe, or very severe [n = 2] or not at all, a little bit, somewhat, quite a bit, or very much [n = 1]) and difficulty terms (no difficulty, slight, some, a lot of, unable [n = 7], or a phrase to describe difficulty [n = 1]). Countries that needed translation (Argentina, China, and Germany) used the validated translation from the face validity studies (stage 3). Items that were added or modified from the face validity study underwent back translation into English alongside modification by an independent translation company. International teams proofread and approved the final versions.

Background information and other health and wellbeing questionnaires were also included to support analysis of domain structure. The additional questionnaires were the EQ-5D 3-level version (EQ-5D-3L)[13] and EQ-5D 5-level version (EQ-5D-5L),[14] the short Warwick-Edinburgh Mental Wellbeing Scale (SWEMWBS),[15] and Adult Social Care Outcomes Toolkit [ASCOT][16]). EQ-5D-3L and EQ-5D-5L are generic health measures with 5 dimensions and 3 or 5 levels of severity, respectively. The SWEMWBS is a 7-item measure covering positive mental wellbeing. ASCOT, a measure of social care related quality of life with 9 items, was only included in the United Kingdom, the United States, and Australia because translated versions were not available. Items from the SWEMWBS and EQ-5D were used to support estimation of latent constructs for the IRT and CFA. Average scores for the measures were used to describe the samples; ASCOT was scored using UK public preferences,[16] SWEMWBS was scored by summing across items,[15] and the EQ-5D measures were scored with the relevant country tariffs where available.

Three versions of the survey were created with a different order of the EQ-HWB candidate items to minimize learning and order effects and further randomized as to whether EQ-5D-3L or EQ-5D-5L appeared last (to support separate analysis on a comparison between these 2 instruments) making 6 different versions. Positively (n = 19) and negatively (n = 36) worded frequency items were grouped together to minimize the number of reversals of the meaning of the response options (eg, whether "often" represents higher or lower wellbeing). Items using "difficulty" response options were also grouped together. All other questionnaires were presented after the EQ-HWB candidate items. The same versions were administered across all countries, with approved EuroQol translations for EQ-5D to German, Argentinian Spanish, and simplified Chinese and relevant translations for the SWEMWBS.[17-19] A single company (Accent) managed the data collection.

All participants provided an informed consent. Participants recruited in the United Kingdom via NHS Trusts, primary care organizations, and other organizations were given a £5 voucher, which was sent on receipt of the questionnaire. Online participants were rewarded based on their specific panel agreements, which was mainly points. Ethical approval was obtained for all the studies.

Analysis

The main aim of the data analysis was to assess item performance from a psychometric perspective to support selection of items for a long measure and a shorter measure, which would be suitable for valuation. The analysis also sought to confirm the domain structure. Classical psychometric analysis was undertaken exploring responses (inconsistencies, missing data, distribution) and sensitivity to known group differences. Factor analysis and IRT were used to assess dimensionality and performance of items. All items were recoded such that a higher score reflected poorer health or wellbeing. An analysis protocol was developed and used to support consistent analysis across the countries with modification based on sample size and groups that were included. A summary of the analysis methods is presented in Table 1[20-26]; further detail is available in the Supplemental Technical Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1361.

To support the consultation process that was used to inform item selection for the 2 EQ-HWB measures,[1] each country team summarized the all the psychometric evidence and face validity evidence[27] using a 4 to 1 scale (performs very well, fairly well, mixed evidence, and performs poorly) for each item. The project criteria for item selection[5] aided item prioritization. This included judgments on whether there was evidence that items were unacceptable (eg, ambiguous, offensive), were interpreted and answered differently for different people (eg, experiencing differential item functioning [DIF]), or were too mild or extreme to be appropriate for inclusion in a generic measure.

Results

Sample

A total of 4879 participants were recruited during 2018 and 2019 to take part in the psychometric survey across the 6 countries including people with long-term conditions, social care users, and carers (Table 2); 49 respondents were dropped because of inconsistencies: United Kingdom (32), Germany (6), China (0), Argentina (11), and the United States (0). Mean age was 46.8 (SD 17.8) and 51.5% were female and 36.2% were carers.

No respondents took up the offer of a face-to-face interview and 15 chose to complete the survey online after receiving a article-based questionnaire. Those who completed online were younger than those who completed the article-based questionnaire.

Missing Items and Distribution

The average missing data for the candidate items were 2.3% from the UK article-based surveys and 0.6% for the United States, which was the only online collection that did not force responses. The proportion of missing data in the UK article survey ranged from 0.6 to 4.5% (see Supplemental Appendix UK Table 2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1361) with the exception of one question ("I felt able to cope with my day-to-day life") that had 19.5% missing because of a problem with the printing of one version of the questionnaire. In the United States, the proportion ranged from 0% to 6.4%, the highest being for the item "I was able to do the things I wanted to do" (see Supplemental Appendix USA Table 2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1361).

A summary of the results of the classical psychometric analysis is presented in Appendix Table A.3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1361. Most of the items did not have high ceiling effects with the exception those from the self-care, mobility, hearing, and seeing subdomains. Only 2 items had <5% reporting at the highest end: "I felt worried" in Argentina and "I felt able to cope with my day-to-day life" in China.

There was evidence of <5% at the lowest level for most items, with slightly less distribution concerns in the UK and Australian data. There were few respondents reporting the worse option for items within the activity domain, items in the safety subdomain, the items for pain and discomfort with severity response options, and the items "I felt calm" and "I got along well with people around me."

Known Group Validity

Most of the items were able to detect known group differences across the physical health conditions with moderate to large effect sizes in most countries. Appendix Table A.3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1361 provides a summary based on the lowest effect size across the identified health conditions (see also Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1361). Most items were able to discriminate well between those with and without an identified mental health condition in all countries (Appendix Table A.3 and Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1361).

In the larger UK and US data sets, comparisons also included severity of mental health condition and severity of arthritis based on self-reported of current mobility and anxiety/depression. For some items, effect sizes for these comparisons were lower, which helped in distinguishing between items in these countries (see Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1361).

The support question "I had support when I needed it" performed poorly across all countries for physical and mental health group differences, despite a similar item "I felt unsupported by other people" having moderate to high effect sizes. For the self-care items, those with frequency response options perform worse than those with difficulty response options. Items aiming to tap into final outcomes (eg, My personal needs were met) which could be attained through respondents' own functioning or through their care provision showed lower effect sizes.

Appendix Table A.3 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1361 also shows effect sizes for high- versus low-hour carers. Effect sizes are moderate to low and do not clearly distinguish between items (effect sizes for having a caring role are shown in the Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1361).

Domain Structure: Exploratory Factor Analysis and CFA

Exploratory factor analysis results showed that 3 factors contained most of the items and combined into a positively worded broad wellbeing/mental health factor, a negatively worded broad wellbeing/mental health factor, and a factor for SWEMWBS items. These results suggested the need to model CFA using a bifactor approach in which a negatively worded measurement factor and a positively worded measurement factor were included alongside

Table 1. Analysis methods.

Assessment area	Measurement property	Analysis
Data cleaning	Inconsistency	- Inconsistency—where the EQ-5D-3L and EQ-5D-5L were both collected (all countries except Australia), reporting top level for a domain in 1 EQ-5D version yet the bottom level in the other EQ-5D version was judged as evidence of a potential lack of concentration and hence inconsistency - Ticking >1 response option (article-based UK data only)
Distribution of responses	Missing data	Missing data (article-based UK data and online USA only because these were the only surveys in which skipping items was possible). Items were flagged if they had >5% missing data.
	Ceiling and floor effects	Distribution across severity levels was assessed and items were judged as potentially problematic if they had skewed distributions such that either >70% or <5% responded in the top or bottom category. Consideration was given to the actual item and population because for some items, such as "seeing," we would expect to see a very skewed distribution.
Validity	Known group	The ability of items to discriminate between groups with known differences was assessed. Cohens D (mean difference divided by pooled SD) was estimated to assess sensitivity based on standard thresholds: 0.2 to <0.5, 0.5 to <0.8, and 0.8 or more denote small, medium, and large effect sizes, respectively.[20] Mann-Whitney-Wilcoxon tests that more appropriately account for the 5-point ordinal scale were also conducted. This is calculated based on the percent of cases in which a random observation from group 0 is higher than a random observation from group 1, plus half the probability that the values are tied. Groups included: - Those with a health condition compared with a healthy group defined as an EQ-VAS score of 80 or above and no long-term condition reported; matched by age group - Mild vs severe mobility impairment judged based on EQ-5D-5L mobility item for those self-reporting arthritis (levels 1-3 vs 4-5) - Mild vs severe mental health condition judged based on EQ-5D-5L depression/anxiety item for those self-reporting depression, anxiety, or another mental health condition (levels 1-3 vs 4-5) - Carers vs noncarers; matched by age group, being female, and presence of long-term condition - Carers with low burden (caring 1-19 hours) vs carers with high burden (>19 hours); matched by age group, being female, and presence of long-term condition
Measurement structure	Correlation	Full correlation matrix using Spearman rank correlations to identify the level of correlation of items within dimensions and across dimensions. Items that had low correlations with other items within their subdomain (<0.5) or high correlations with items from other subdomains (>0.7) were flagged because this indicated they may be measuring other constructs.
	Multidimensional relationship	- EFA was used to further understand the correlation patterns within the data and to determine whether a bifactor model[21] was required. Bifactor models are used in CFA where items reflect >1 construct.[22] - CFA was used to test the working conceptual model. The CFA excludes hearing, seeing, discomfort, and sleep because these only had 1 or 2 items and they had been found to be independent in separate analyses (see Technical Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1361). A bifactor CFA model was used to reflect both the conceptual framework and results from the correlation analysis and EFA. Highly correlated items (>0.8) from different subdomains were analyzed together. Items from the EQ-5D and SWEMWBS were included in the CFA where necessary to aid model identification and robustness. All models were undertaken using UK data and then tested on the other country data. CFA models were judged based on the following fit indices: RMSEA <0.6 taken as good, CFI >0.95 taken as good, and TLI >0.95 taken as good.[23] To assess the bifactor results, items were flagged if the ratio of the loading to factor of interest (a measure of association between the item and the construct we are trying to measure) was relatively low compared with loading onto the measurement factor. Given that no cutoff recommendation could be found in the literature, we judged that a value of below 1.75 times was potentially problematic.
	Multi-item relationship	The graded response IRT model was used to assess the fit and performance of items within each factor representing a domain from the CFA. Analysis was done separately for each domain. - Item fit was confirmed using S_X^2.[24] We applied a threshold of P<.01 as a potential flag for misfit; this is lower than the standard 0.05 because of multiple testing. - Local independence was tested by assessing residual correlation between items within the domain. Items were flagged if they had a residual correlation > 0.25.[25]

continued on next page

Table 1. Continued

Assessment area	Measurement property	Analysis
	Item performance	- The discrimination (slope) of each item indicates how well the item differentiates between subjects with different levels of the latent factor. A higher value is preferred. Item discrimination can be classified as follows: 0 is "none"; 0.01-0.34 is "very low"; 0.35-0.64 is "low"; 0.65-1.34 is "moderate"; 1.35-1.69 is "high"; ≥ 1.7 is "very high."[26] - Item threshold parameters (difficulty) of each item show the level of the latent factor above which the probability of choosing that response option or higher is >50%. - CCCs show the probability of choosing a category for each value of the latent variable (theta). The CCCs were studied for each item to evaluate the function of each response category, in particular whether there is a specific range of the latent factor where that category is most likely to be chosen. Items with nonoptimal response categories were flagged. - Item information functions show how well and precisely each item measures the latent factor at different levels of the latent factor, using a measure called "information." Preferred items are those that provide high level of information across the full range of the latent factor score. - DIF tests assesses whether different groups treat the questions in the same way. DIF tests were based on age (young [18-44 years] vs old [+45 years]), gender, having a degree, mental health condition, and being a carer. Items were flagged if they exhibited DIF. DIF analyses were performed within the IRT model. These were based on a likelihood ratio test of main vs nested models that constrained the discrimination and difficulty parameters to be the same across the groups of interest. Due to the large number of DIF tests (5 groups across all candidate items for each country), a $P<.01$ was used.

CCC indicates category characteristic curve; CFA, confirmatory factor analysis; CFI, comparative fit index; DIF, differential item functioning; EFA, Exploratory factor analysis; EQ-5D-3L, EQ-5D 3 level version; EQ-5D-5L, EQ-5D 5 level version; EQ-VAS, EQ-visual analog scale; IRT indicates item response theory; RMSEA, root mean square error of approximation; SWEMWBS, short Warwick-Edinburgh Mental Wellbeing Scale; TLI, Tucker-Lewis index; UK, United Kingdom; USA, United States of America.

the subdomains (both of which were constrained to have zero correlation with the subdomain factors).

The initial CFA model was established using UK data from a combination of the original conceptual model, studying the correlation matrix with particular attention to where items did not appear to fit well within their subdomain, and the results from the exploratory factor analysis. The inclusion of measurement factors via a bifactor model improved model performance although most of the variance in items (>69%) was explained by the domain factor rather than the measurement factors.

The initial analysis found that the domain daily activity was not well identified. This subdomain was originally included in the conceptual model to cover usual daily activity such as work, travel shopping, and housework. Three potential items designed to pick up this domain (EQ-5D usual activity, I could do what was needed, I was able to do daily activities) were not highly correlated. Consequently, these items were moved to domains where they showed higher correlation patterns. The item "I could do what I needed" was moved into meaningful/valuable activity, and the item "I was able to do daily activities" was moved into mobility.

The item "I was able to do the things I wanted to do" with severity response choices was originally intended to be within the meaningful/valuable activity subdomain yet correlated <0.75 with other items in the domain and correlated >0.75 with other items from the control domain. Therefore, it was moved into the control subdomain.

A well-fitting bifactor CFA model was attained for the UK data (Fig. 1) in which there were separate factors for subdomains in the feelings and emotions domain (happy, hope, anxiety, safety, and anger), activity domain (self-care, enjoyable activity, mobility), autonomy domain (control and coping), and physical sensations domain (pain and energy). Self-respect only had one subdomain. Relationships subdomains did not separate out into different factors whereas the cognition subdomains did not have enough items to separate them out. The positively and negatively worded measurement factors were also included.

The CFA factors were highly correlated yet merging those factors that correlated most highly (0.961 between anxiety and coping) resulted in reduced model fit. The UK model also achieved good fit on the Australian and US data. The model required some minor modification to fit the German data where good model fit was attained when energy (which included only 2 items) was dropped and mobility and self-care factors were combined. Similarly, for Argentina, it also required combining self-worth and coping. The model did not fit the China data as well; many of the feelings subdomains needed to be combined. The data from China also did not fit the positive and negative measurement factors in the same manner.

Item Performance: IRT Results

The IRT analysis found that response categories worked well for most items (see Appendix Table A.4 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1361). Response categories for the self-care items using frequency response options did not work well in any of the countries. The response categories for items "I had support when I needed it," "I got along well with people around me," "I felt calm," "I felt like a failure," and "I felt unsafe" had evidence of merged categories in at least 1 country. These problems may indicate that fewer levels are required for the response options for these items.

Evidence from the IRT models (including item fit [from the S statistic]), the influence of the positive/negative measurement factors in the CFA, and the amount of information each item added and where this occurred on the latent construct (from the item information functions [IIFs]) was interpreted in the light of whether items are conceptually aiming to measure something

Table 2. Summary statistics from the psychometric surveys

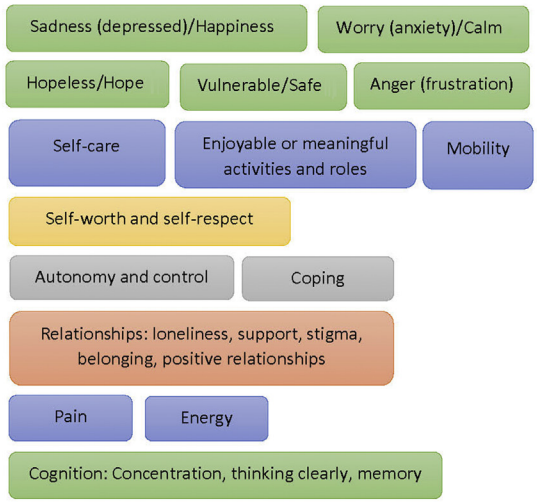
Respondent characteristics	UK (online), n (%)	UK (article), n (%)	Argentina, n (%)	Australia, n (%)	China, n (%)	Germany, n (%)	USA, n (%)
Age, mean (SD)	41.4 (15.5)	63.5 (16.1)	37.4 (12.8)	49.9 (16.9)	35.9 (8.6)	44.8 (17.1)	53.8 (17.5)
Female	752 (58.0)	345 (55.0)	201 (40.4)	201 (39.1)	298 (60.0)	253 (51.0)	436 (48.3)
Degree	660 (50.9)	229 (36.5)	267 (53.7)	242 (47.1)	378 (76.1)	250 (50.4)	398 (44.1)
Carer	393 (30.3)	198 (31.6)	339 (68.2)	115 (22.4)	226 (45.5)	280 (56.5)	196 (21.9)
Hours cared							
1-19	151 (38.4)	55 (26.7)	145 (45.2)	56 (48.7)	71 (31.4)	123 (43.9)	92 (50.3)
20-49	132 (33.6)	36 (17.5)	110 (34.3)	24 (20.9)	134 (59.2)	93 (33.2)	48 (26.2)
≥50	86 (21.9)	100 (48.5)	66 (20.6)	28 (24.3)	17 (7.5)	47 (16.8)	43 (23.5)
Social care	226 (17.4)	131 (20.9)	287 (57.8)	77 (15.0)	NA	131 (26.4)	86 (9.5)
Long-term condition	906 (69.9)	554 (88.4)	317 (63.8)	374 (72.8)	357 (71.8)	333 (67.1)	NA
Asthma	211 (16.3)	106 (16.9)	68 (13.7)	78 (15.2)	44 (8.9)	40 (8.1)	76 (8.4)
Arthritis	193 (14.9)	166 (26.5)	36 (7.2)	100 (19.5)	45 (9.1)	25 (5.0)	161 (17.8)
Heart conditions	86 (6.6)	110 (17.5)	62 (12.5)	41 (8)	43 (8.7)	35 (7.1)	86 (9.5)
Stroke	25 (1.9)	35 (5.6)	6 (1.2)	10 (1.9)	12 (2.4)	14 (2.8)	19 (2.1)
Overactive thyroid	19 (1.5)	8 (1.3)	18 (3.6)	12 (2.3)	25 (5.0)	14 (2.8)	14 (1.6)
Underactive thyroid	60 (4.6)	55 (8.8)	28 (5.6)	12 (2.3)	18 (3.6)	32 (6.5)	68 (7.5)
Bronchitis/emphysema	32 (2.5)	18 (2.9)	18 (3.6)	31 (6)	40 (8.0)	28 (5.6)	42 (4.8)
Liver conditions	23 (1.8)	14 (2.2)	14 (2.8)	14 (2.7)	29 (5.8)	3 (0.6)	23 (2.5)
Cancer	76 (5.8)	62 (9.9)	95 (19.1)	19 (3.7)	13 (2.6)	196 (39.5)	NA
Diabetes	136 (10.5)	150 (23.9)	145 (29.2)	50 (9.7)	71 (14.3)	59 (11.9)	122 (13.5)
Epilepsy	34 (2.6)	13 (2.1)	17 (3.4)	4 (0.8)	4 (0.8)	11 (2.2)	4 (0.4)
High blood pressure	162 (12.5)	163 (26.0)	63 (12.7)	107 (20.8)	81 (16.3)	92 (18.6)	276 (30.6)
IBS	173 (13.3)	62 (9.9)	41 (8.3)	33 (6.4)	11 (2.2)	27 (5.4)	49 (5.4)
Depression	245 (18.9)	84 (13.4)	83 (16.7)	159 (30.9)	106 (21.3)	62 (12.5)	113 (12.5)
Generalized anxiety disorder	230 (17.7)	67 (10.7)	128 (25.8)	144 (28)	89 (17.9)	36 (7.3)	98 (10.9)
Back pain	NA	NA	NA	157 (30.5)	NA	NA	NA
Disability	NA	NA	NA	47 (9.1)	19 (3.8)	NA	102 (11.3)
Chronic hepatitis B	NA	NA	NA	NA	120 (24.1)	NA	NA
HIV/AIDS	NA	NA	NA	NA	101 (20.3)	NA	6 (0.7)
Other physical	265 (20.4)	144 (23.0)	30 (6.0)	94 (18.3)	8 (1.6)	57 (11.5)	192 (21.3)
Other mental	141 (10.9)	83 (13.2)	4 (0.8)	35 (6.8)	2 (0.4)	25 (5.0)	52 (5.8)
Total	1296	627	497	514	497	496	903
EQ-5D-3L, mean (SD)	0.646 (0.34)	0.659 (0.32)	0.573 (0.31)	NA	0.821 (0.17)	0.694 (0.24)	0.79 (0.19)
EQ-5D-5L, mean (SD)	0.737 (0.26)	0.74 (0.26)	0.879 (0.14)	0.733 (0.23)	0.820 (0.18)	0.773 (0.27)	0.76 (0.26)
ASCOT, mean (SD)	0.725 (.26)	0.802 (.22)	NA	0.764 (0.24)	NA	NA	0.84 (0.19)
SWEMWBS, mean (SD)	22 (7)	25.7 (6.8)	24.2 (6)	21.8 (6.1)	25.9 (5.8)	24.9 (6.0)	25.75 (6.74)

Note. Detail regarding EQ-5D utility scoring for each country is shown in Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1361. Note that respondents may report having >1 health condition.
ASCOT indicates Adult Social Care Outcomes Toolkit; HIV/AIDS, human immunodeficiency virus/acquired immunodeficiency syndrome; IBS, irritable bowel syndrome; NA, not available; SWEMWBS, short Warwick-Edinburgh Mental Wellbeing Scale; UK, United Kingdom; USA, United States of America.

slightly different to the other items within the construct and the positive/negative framing of items within the domain (see Supplemental Material found at https://doi.org/10.1016/j.jval.2021.11.1361). For example, for the anxiety domain, the IIF for the item "I felt anxious" provided more information than the item "I felt worried" in the United States, United Kingdom, Germany, and Australia and a similar level to other items in China and Argentina, suggesting a stronger performance for the former item. At the same time, the IIFs for the other item in this domain, "I felt calm," was well below other IIFs in all countries, but this was interpreted in the light of this being a positively rather than negatively framed item.

There is some evidence of DIF for many of the items (see Appendix Table A.4 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1361), most commonly age-related DIF. Some of the items in the domains of self-care, happiness, and self-worth were found to have relatively higher occurrences of DIF. DIF was also identified in items in the control and anxiety

Figure 1. Confirmatory factor analysis (excluding the domains seeing, hearing, sleep problems, and discomfort).



	RMSEA (95% CI)	CFI	TLI
UK	0.059 (0.059 0.060)	0.958	0.953
Argentina: UK model with *Mobility* merged with *Self-care*, *Energy* removed and *Self-worth* merged with *Coping*	0.049 (0.047 0.051)	0.956	0.952
Australia: UK model has good fit	0.057 (0.055 0.059)	0.965	0.961
China: UK model with -*Mobility* merged with *Self-care*. -*Anxiety, Happiness, Hope, Coping, Self-worth* all combined to one factor -*Energy* removed Some items did not fit either the positive or the negative measurement factor	0.072 (0.070 0.074)	0.907	0.901
Germany: UK model with *Mobility* merged with *Self-care*, and *Energy* removed.	0.052 (0.050 0.054)	0.961	0.957
USA: UK model has good fit	0.055 (0.054 0.057)	0.969	0.966

CFI indicates comparative fit index; CI, confidence interval; RMSEA, root mean square error of approximation; TLI, Tucker-Lewis index; UK, United Kingdom.

domains for those in the sample reporting a mental health condition.

Item Performance Overall

Based on the psychometric evidence, approximately half of the items (n = 32) were judged to perform very or fairly well by at least 5 countries. Some had mixed evidence (n = 27) whereas a limited number of items (n = 7) were judged as performing poorly by any country. This evidence was used to support the consultation with wider stakeholders.[1]

Discussion

Overview

The results from testing the candidate items indicated that most items performed well across the patient groups. Missing data did not discriminate between items, in part because most data were collected via an online platform where skipping items was not possible. Most items achieved a good spread across the response choices. Skewed distributions were present for items in the domains of mobility, self-care, hearing, seeing, and safety reflecting expectations that most respondents would not have problems in these domains and only very few respondents would have severe problems. Pain and discomfort measured by severity also had very few respondents using the poorest response option; nevertheless, identifying patients with very severe pain can be important when evaluating interventions.

Most items were able to distinguish between respondents with physical and mental health conditions and by severity of condition where this was tested; hence, this provides little basis for discrimination. The known group validity evidence was mixed for carers with mostly small or insignificant effect sizes for carers versus noncarers and across high- versus low-hour carers. Although we may have expected to see high-hour carers scoring lower on some items (such as relationships and activities) without more detail on the type of caring, it is hard to draw conclusions from this. Furthermore, the matching across carers based on age, gender, and long-term condition may not have adequately captured other related characteristics.

The conceptual model was generally confirmed although there was evidence of high correlation between the final factors. The data were best modeled as a bifactor model with positive and negative measurement factors along with the construct/domain factors. Although we use the terms "negatively and positively worded measurement factors," we acknowledge that we do not have a clear understanding of what is behind these latent constructs. The CFA identified a well-fitting model with 15 domains for UK, Australian, and US data. A further 4 domains (seeing, hearing, discomfort, and sleep problems) from the original conceptual model were not included for testing in this CFA but had been identified as independent in earlier (secondary) data analysis (see Technical Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1361). Achieving well-fitting model for other countries involved removing energy and merging of domains leaving 13 separate domains for Germany, 12 for Argentina, and 10 for China. The items designed to capture meaningful and valuable activity, problems with daily activities, and feelings of control and autonomy did not clearly identify these different constructs.

The CFA model relied upon controlling for 2 measurement factors; nevertheless, there is no well-established method for conducting IRT on multidimensional models. Given that most (≥69% for all domains) of the variance in items in the UK data was explained by domain factors rather than the measurement factors, it was reasonable to conduct the IRT on separate domains.

Strengths and Limitations

The psychometric analysis relied on large mixed samples in different countries with different languages and cultural values. A mix of patients with physical and mental health conditions and social care users and carers were targeted in the different countries, which enabled assessment of the questions in future users. Online recruitment enabled certain patient groups to be recruited in a timely and cost-effective way that was standardized across the countries. Accepted methods for assessing the psychometric

performance of items were applied to inform the selection of items for the measure. Focusing on specific analysis (eg, excluding factor analysis for domains such as hearing and seeing) and using available data (eg, using items from other questionnaires to allow factor analysis to be undertaken) ensured that there was a balance between research requirements and respondent burden.

Nevertheless, there were some limitations. There were few respondents in the lowest levels of pain/discomfort that may be important in assessments of interventions but those in very severe pain may be harder to recruit. Although social care users and carers were included, there was some ambiguity with regard to impact on health and wellbeing and no clear markers to test sensitivity. Receiving social care has an ambiguous interpretation on quality of life: it may indicate a greater need for social care or the effectiveness of receiving services; therefore, known group assessment was not undertaken for this group. For the caring role, caring for a friend or relative may suggest a person has higher wellbeing (such as close relationships[28]); alternatively, the caring role may reduce health and wellbeing. The caring burden was proxied here through hours of caring, which may be a weak indicator.

Data were drawn from 6 countries but these do not cover the same samples, which limits our ability to make between country comparisons. Online recruitment meant assessment of missing data could not be fully undertaken although this was partly mitigated by the inclusion of a article-based survey in the United Kingdom. In this study, online respondents were younger than those who completed the measure by article. Practicalities of the article version distribution resulted in a limited use of randomization of the order of items, which left a risk of order effects.

The known group analysis was mainly patient versus healthy population comparison. This is only a crude test and may not discriminate between respondents (indeed we find high effect sizes for almost all items). Although there were comparisons of severity, these were based on relevant EQ-5D dimension levels for arthritis and mental health and there were no clinical severity indicators. The aim for the final measure is to be sensitive to changes in health and wellbeing from treatment or services provided to meet individual needs. The known group difference assessment did not include indicators that could assess this type of sensitivity.

There was some evidence of "negative" and "positive" measurement factors within the factor analysis suggesting that some respondents answer negatively and positively framed questions differently.

Although the psychometric analysis here suggests the potential to merge subdomains, particularly for China, Germany, and Argentina, where the best fitting model arose when subdomains were merged, this merging is data driven and could arise from a common co-occurrence of problems across conceptually distinct domains. This is a particular concern in the non-UK samples, which had a more limited diversity of patient groups. The extent to which conceptually separate domains could be treated as merged for future patient groups or users of the measure is not clear.

The initial conceptual model[3] was developed from qualitative evidence from Western context. This may explain why the model achieves a poorer fit within the data from China and points to the difficulty in developing a generic measure valid in all countries and cultures.

The DIF analysis relied upon the significance of DIF and did not explore the magnitude of DIF; hence, it is unclear how problematic the identified DIF is. Additionally, some DIF analysis comparing subgroups with smaller sample sizes may have lacked power.[29]

The focus on this analysis was on providing evidence to support item selection that would complement qualitative face validity work. Further psychometric analysis on the final EQ-HWB instrument could address many other interesting questions.

Conclusions

A candidate pool of EQ-HWB items (n = 64) was tested to assess their psychometric performance to support the selection of items for inclusion in the new measures.[1] The international evidence found that most of the items performed well but highlighted that some items had mixed evidence and may perform less well in different contexts. The conceptual model indicated that there were 19 subdomains that reflect health and wellbeing in this diverse population.

Psychometrics is an essential step to select items to generate a valid measure of health and wellbeing. This analysis combined multiple methods to provide a broad range of evidence for this purpose.

Supplemental Materials

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2021.11.1361.

Article and Author Information

Accepted for Publication: November 4, 2021

Published Online: January 13, 2022

doi: https://doi.org/10.1016/j.jval.2021.11.1361

Author Affiliations: Melbourne School of Population and Global Health, University of Melbourne, Victoria, Australia (Peasgood); School of Health and Related Research, University of Sheffield, Sheffield, England, UK (Peasgood, Mukuria, Brazier); Department of Health Economics and Health Care Management, School of Public Health, Bielefeld University, Bielefeld, Germany (Marten, Kreimeier, Greiner); Saw Swee Hock School of Public Health, National University of Singapore, Singapore (Luo); Centre for Health Economics Research and Evaluation, University of Technology Sydney, New South Wales, Australia (Mulhern); Department of Pharmacy Systems, Outcomes and Policy, College of Pharmacy, University of Illinois Chicago, Chicago, IL, USA (Pickard, Monteiro, Kuharic); Institute for Clinical Effectiveness and Health Policy, Buenos Aires, Argentina (Augustovski, Gibbons, Belizan); Deakin Health Economics, School of Health and Social Development, Deakin University, Geelong, Australia (Engel); Health Services Management Department, Guizhou Medical University, Guiyang, China (Yang); QualityMetric Incorporated, LLC, Johnston, RI, USA (Bjørner).

Correspondence: Tessa Peasgood, PhD, Centre for Health Policy, Melbourne School of Population and Global Health, University of Melbourne, Level 4, 207 Bouverie St, Melbourne, Victoria, Australia 3010. Email: Tessa.Peasgood@unimelb.edu.au

Author Contributions: *Concept and design*: Peasgood, Mukuria, Brazier, Pickard, Engel
Acquisition of data: Peasgood, Mukuria, Marten, Kreimeier, Luo, Mulhern, Greiner, Pickard, Augustovski, Engel, Yang, Monteiro, Kuharic, Belizan
Analysis and interpretation of data: Peasgood, Mukuria, Brazier, Marten, Kreimeier, Mulhern, Greiner, Pickard, Augustovski, Engel, Gibbons, Yang, Monteiro, Kuharic, Belizan, Bjørner
Drafting of the manuscript: Peasgood, Mukuria, Brazier, Engel, Mulhern
Critical revision of the paper for important intellectual content: Peasgood, Mukuria, Brazier, Marten, Kreimeier, Luo, Mulhern, Greiner, Pickard, Augustovski, Engel, Gibbons, Yang, Monteiro, Kuharic, Belizan, Bjørner
Statistical analysis: Peasgood, Mukuria, Mulhern, Augustovski, Engel, Gibbons, Monteiro, Kuharic, Bjørner
Provision of study materials or patients: Peasgood, Marten, Kreimeier, Greiner, Augustovski

Obtaining funding: Peagood, Mukuria, Brazier, Marten, Kreimeier, Luo, Mulhern, Greiner, Pickard, Engel, Yang
Administrative, technical, or logistic support: Peasgood, Marten, Kreimeier, Greiner, Engel, Monteiro, Kuharic
Supervision: Pickard

Conflict of Interest Disclosures: Drs Peasgood, Brazier, Mulhern, Engel, and Yang and Ms Belizan reported receiving grants from the Medical Research Council and the EuroQol Research Foundation during the conduct of this study. Dr Mukuria reported receiving grants from the EuroQol Research Foundation during the conduct of this study and outside the submitted work and reported being a member of the EuroQol Research Association. Dr Brazier, Marten, Kreimeier, Mulhern, Greiner, Engel, and Yang reported being members of the EuroQol Group. Dr Brazier reported being a past member of the EuroQol Group Executive, reported receiving grants and personal fees from the EuroQol Research Foundation outside the submitted work, and reported receiving royalties paid to the University of Sheffield for the use of the SF-6D preference-based measure of health outside the submitted work. Drs Marten, Kreimeier, and Greiner reported receiving grants and nonfinancial support from the EuroQol Research Foundation during the conduct of this study. Dr Luo reported receiving grants and personal fees from EuroQol Research Foundation during the conduct of the study and outside the submitted work. Drs Luo and Mulhern are editors for *Value in Health* and had no role in the peer-review process of this article. Drs Pickard, Augustovski, and Gibbons and Mses Monteiro and Kuharic reported receiving grants from the EuroQol Research Foundation during the conduct of the study. Ms Kuharic reported receiving a fellowship from Takeda Pharmaceuticals USA outside the submitted work. No other disclosures were reported.

Funding/Support: This work was supported by grant 170620 from the UK Medical Research Council and grants 20180460, 20180600, 20190260, 20180450, 20180580, and 20180520 from the EuroQol Research Foundation.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgment: The authors acknowledge the support of the National Institute for Health Research (NIHR) Yorkshire and Humber Applied Research Collaboration (formerly CLAHRC) and the NIHR Clinical Research Network (CRN). The authors acknowledge the invaluable contributions of members of the project steering group, advisory group, and public and patient involvement and engagement groups and Julie Johnson for project administration. We also thank members of staff at the National Institute for Health and Care Excellence who attended project workshops and gave valuable feedback. The authors also thank members of the EuroQol Group Association for their input at plenary and academy meetings and the EuroQol office for their support. The authors also thank the NIHR CRN and the South West Peninsula, Northwest Coast, North Thames, and North East and North Cumbria CRNs, in addition to National Health Service organizations across primary and secondary care for their support with recruitment. Finally, The authors acknowledge the contribution of all the patients, social care users, informal carers, and other members of the public who took part in all the studies across the different countries.

REFERENCES

1. Brazier J, Peasgood T, Mukuria C, et al. The EQ-HWB: Overview of the development of a measure of health and well- being and key results. *Value Health*. 2022, in press.
2. Peasgood T, Mukuria C, Carlton J, et al. What is the best approach to adopt for identifying the domains for a new measure of health, social care and carer-related quality of life to measure quality-adjusted life years? Application to the development of the EQ-HWB? *Eur J Health Econ*. 2021;22:1067–1081.
3. Mukuria C, et al. A targeted review of qualitative evidence on domains of quality of life important for patients, social care users and informal carers to inform the development of the EQ health and wellbeing (EQ-HWB).
4. Carlton J, PT, Mukuria C, Connell J, et al. Generation, selection and face validation of items for a new generic measure of quality of life, the EQ health and wellbeing (EQ-HWB). *Value Health*. 2022 (in press). doi 10.1016/j.jval.2021.12.007
5. Peasgood T, Mukuria C, Carlton J, Connell J, Brazier J. Criteria for item selection for a preference-based measure for use in economic evaluation. *Qual Life Res*. 2021;30(5):1425–1432.
6. Fayers PM, Machin D. *Quality of Life: The Assessment, Analysis and Interpretation of Patient-Reported Outcomes*. Chichester, United Kingdom: John Wiley & Sons; 2013.
7. Cappelleri JC, Jason Lundy J, Hays RD. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clin Ther*. 2014;36(5):648–662.
8. Qual outcomes framework (QOF): 2015-2016. NHS Digital. https://digital.nhs.uk/data-and-information/publications/statistical/quality-and-outcomes-framework-achievement-prevalence-and-exceptions-data/quality-and-outcomes-framework-qof-2015-16. Accessed February 26, 2018.
9. Comrey A, Lee H. *A First Course in Factor Analysis*. Hillsdale, NJ: Lawrence Eribaum Associates. Inc, Publishers; 1992.
10. Netemeyer RG, Bearden WO, Sharma S. *Scaling Procedures: Issues and Applications*. Thousand Oaks, CA: Sage Publications; 2003.
11. Tennant R, Hiller L, Fishwick R, et al. The Warwick-Edinburgh mental well-being scale (WEMWBS): development and UK validation. *Health Qual Life Outcomes*. 2007;5(1):1–13.
12. Richardson J, Iezzi A, Khan MA, Maxwell A. Validity and reliability of the Assessment of Quality of Life (AQoL)-8D multi-attribute utility instrument. *Patient Patient Centered Outcomes Res*. 2014;7(1):85–96.
13. Brooks R, Group E. EuroQol: the current state of play. *Health Policy*. 1996;37(1):53–72.
14. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011;20(10):1727–1736.
15. Stewart-Brown S, Tennant A, Tennant R, Platt S, Parkinson J, Weich S. Internal construct validity of the Warwick-Edinburgh mental well-being scale (WEMWBS): a Rasch analysis using data from the Scottish health education population survey. *Health Qual Life Outcomes*. 2009;7(1):1–8.
16. Netten A, Burge P, Malley J, et al. Outcomes of social care for adults: developing a preference-weighted measure. *Health Technol Assess*. 2012;16(16):1–166.
17. Lang G, Bachinger A. Validation of the German Warwick–Edinburgh mental well-being scale (WEMWBS) in a community-based sample of adults in Austria: a bi-factor modelling approach. *J Public Health*. 2017;25(2):135–146.
18. Dong A, Chen X, Zhu L, et al. Translation and validation of a Chinese version of the Warwick–Edinburgh mental well-being scale with undergraduate nursing trainees. *J Psychiatr Ment Health Nurs*. 2016;23(9-10):554–560.
19. Serrani Azcurra D. Translation, Spanish adaptation and validation of the Warwick-Edinburgh Well-being Scale in a sample of Argentine older adults. *Acta Colomb Psicol*. 2015;18(1):79–93.
20. Cohen J. A power primer. *Psychol Bull*. 1992;112(1):155–159.
21. Reise SP, Bonifay WE, Haviland MG. Scoring and modeling psychological measures in the presence of multidimensionality. *J Pers Assess*. 2013;95(2):129–140.
22. Böhnke JR, Croudace TJ. Calibrating well-being, quality of life and common mental disorder items: psychometric epidemiology in public mental health research. *Br J Psychiatry*. 2016;209(2):162–168.
23. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equ Model Multidiscip J*. 1999;6(1):155.
24. Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. *Appl Psychol Meas*. 2000;24(1):50–64.
25. Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). *J Clin Epidemiol*. 2008;61(1):17–33.
26. Baker FB. The basics of item response theory. Second edition. ERIC. https://eric.ed.gov/?id=ED458219.
27. Carlton J, et al. Patient and Public Involvement and Engagement (PPIE) Within the Development of the EQ Health and Wellbeing (EQ-HWB). *Value Health*. 2022. In press.
28. Li Q, Loke AY. The positive aspects of caregiving for cancer patients: a critical review of the literature and directions for future research. *Psychooncology*. 2013;22(11):2399–2407.
29. Belzak WCM. Testing differential item functioning in small samples. *Multivariate Behav Res*. 2020;55(5):722–747.