



This is a repository copy of *Generation, selection, and face validation of items for a new generic measure of quality of life : the EQ Health and Wellbeing*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/184736/>

Version: Published Version

Article:

Carlton, J. orcid.org/0000-0002-9373-7663, Peasgood, T., Mukuria, C. orcid.org/0000-0003-4318-1481 et al. (15 more authors) (2022) Generation, selection, and face validation of items for a new generic measure of quality of life : the EQ Health and Wellbeing. *Value in Health*, 25 (4). pp. 512-524. ISSN 1098-3015

<https://doi.org/10.1016/j.jval.2021.12.007>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>



ScienceDirect

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

Themed Section: EQ-HWB

Generation, Selection, and Face Validation of Items for a New Generic Measure of Quality of Life: The EQ-HWB



Jill Carlton, PhD, Tessa Peasgood, PhD, Clara Mukuria, PhD, Janice Connell, BSc, John Brazier, PhD, Kristina Ludwig, PhD, Ole Marten, MSc, Simone Kreimeier, PhD, Lidia Engel, PhD, Maria Belizán, MSc, Zhihao Yang, PhD, Andrea Monteiro, MSc, Maja Kuharic, MSc, Nan Luo, PhD, Brendan Mulhern, PhD, Wolfgang Greiner, PhD, Simon Pickard, PhD, Federico Augustovski, PhD

ABSTRACT

Objectives: This article aims to describe the generation and selection of items (stage 2) and face validation (stage 3) of a large international (multilingual) project to develop a new generic measure, the EQ-HWB (EQ Health and Wellbeing), for use in economic evaluation across health, social care, and public health to estimate quality-adjusted life-years.

Methods: Items from commonly used generic, carer, social care, and mental health quality of life measures were mapped onto domains or subdomains identified from a literature review. Potential terms and items were reviewed and refined to ensure coverage of the construct of the domains/subdomain (stage 2). Input on the potential item pool, response options, and recall period was sought from 3 key stakeholder groups. The pool of candidate items was tested in qualitative interviews with potential future users in an international face validation study (stage 3).

Results: Stage 2 resulted in the generation of 687 items. Predetermined selection criteria were applied by the research team resulting in 598 items being dropped, leaving 89 items that were reviewed by key stakeholder groups. Face validation (stage 3) tested 97 draft items and 4 response scales. A total of 47 items were retained and 14 were modified, whereas 3 were added to the candidate pool of items. This resulted in a 64-item set.

Conclusions: This international multiculture, multilingual study with a common methodology identified many items that performed well across all countries. These were taken to the psychometric testing along with modified and new items for the EQ-HWB.

Keywords: EQ-HWB, face validity, interviews, item selection, quality of life.

VALUE HEALTH. 2022; 25(4):512–524

Introduction

The development of new measures requires several stages to identify the relevant domains and items and further stages to test the validity of items in relevant populations. This includes the key assessments of content validity (how well items reflect the scope of what the questionnaire is trying to measure¹) and face validity (how appropriate, relevant, and understandable items and their response options are).^{2,3} There has been increasing demand for detailed accounts of the steps undertaken during these early stages of developing measures.⁴ This article aims to describe the generation and selection of items (stage 2) and face validation (stage 3) of a large international (multilingual) project to develop a new generic measure, the EQ-HWB, that can be used in economic evaluation across health, social care, and public health to estimate quality-adjusted life-years (QALYs). Brazier et al⁵ fully outline the rationale and the theoretical approach for the

EQ-HWB. It is recognized that measuring health alone ignores that many conditions affect outcomes beyond health.⁶ Such measures have limited ability in capturing outcomes in social care, nor do they take into account the impact of conditions upon informal carers. The use of a single measure will allow for comparison of interventions that affect individuals across sectors and avoid risk of double counting. Having a common measure that is suitable for use across health, social care, and public health will provide better evidence to help support cross-sector decision making.⁷ The EQ-HWB has been developed for adults. Potential future work will explore the suitability of the measure for proxy reporting and child-user versions.

The project encompassed 5 stages outlined in [Figure 1](#). This included (stage 1) a literature review to identify potential domains, (stage 2) item generation, (stage 3) cognitive debriefing to test the face validity of potential items, and (stage 4) psychometric analysis of an article and online survey of potential items. After

this stage, a broad consultation exercise identified items to be included in a long version of the measure (25 items) and a shorter version (9 items) of the EQ-HWB measure. Stage 5 was the valuation phase, where selected items are valued by members of the public (to obtain utility weights for use in the estimation of QALYs). More information on the overview of development of the measure and previous and subsequent stages are reported elsewhere.^{5,8,9}

Methods

A large qualitative review was undertaken in stage 1 that identified 7 themes (feelings and emotions, cognition, activity, self-identity, relationships and social connections, “coping, autonomy, and control,” and physical sensations) with 32 subthemes as important domains and subdomains of the quality of life (QoL) of patients, social care users, and informal carers.⁸ A candidate pool of items was generated for the domains/subdomains (stage 2), and these were then tested with potential future users in an international face validation study (stage 3).

The focus for the overall project was on different populations of health, social care, and informal carers with specific emphasis on using the new measure for economic evaluation. Therefore, it required specific considerations in the context of item generation and face validation to ensure that items were fit for purpose.^{10,11} The criteria that an item was required to meet drew on existing published criteria,^{12,13} which was adapted after consultation with the steering and advisory groups of this project to meet the specific needs of the project in creating a generic health, social care, and carer-related QoL preference-based measure.¹¹

Stage 2: Generation of Candidate Items

Stage 2 drew from the qualitative literature review themes and subthemes in stage 1.⁸ There were 4 steps: (1) sourcing items to map to the 32 subdomains (7 domains); (2) refinement and modification of items; (3) review of items from stakeholder, advisory, and patient and public involvement and engagement (PPIE) groups; and (4) further refinement of items and response options.

Step 2a: Sourcing items to map to domains/subdomains

Concepts and terms from the literature review, categorized in domains and subdomains, were summarized, and possible items were identified from existing questionnaires and item banks. Items from commonly used generic, carer, social care, and mental health QoL measures were mapped onto the domains/subdomains. Information on the source, relevant subdomains, original item wording, alternative wording, response options, and notes on whether there were potential problems with the item based on the criteria, such as covering > 1 concept, were documented.

Step 2b: Refinement and modification of items

Potential terms and items were reviewed by the research team to ensure coverage of the construct of the domains/subdomain. Due to the potentially vast number of existing published items on health and QoL, application of the selection criteria began at early screening stages of item generation. Alternative wording was used to modify items (based on team discussions and consensus) where the original item did not fit the proposed structure or criteria for item selection of the new measure.

Step 2c: Review of items from stakeholder, advisory, and PPIE groups

Input on the potential item pool, response options, and recall period was sought from 3 key stakeholder groups. The project PPIE group participated in a focus group session where they were asked to share their thoughts on each item. A second focus group was held with members of the National Institute for Health and Care Excellence (NICE) Citizens Council who are members of the public including patients and social care users. Two researchers with experience in focus group methods facilitated the focus groups. The project international advisory group (consisting of industry, academics, and developers of measures) also provided comments on the proposed item pool via an online survey. In the survey, background information was provided via a video and report, before participants were asked to highlight problematic items with reasons and to provide alternatives. The potential pool of items was also presented to NICE staff who were asked to provide feedback.

Step 2d: Refinement of items and response options

Findings from step 2c were summarized in a spreadsheet and used to refine item wording (where appropriate) and reduce the number of items within the item pool to take forward into stage 3. This included changing any ambiguous words, adding explanations, and dropping any items that were considered particularly problematic based on the feedback received.

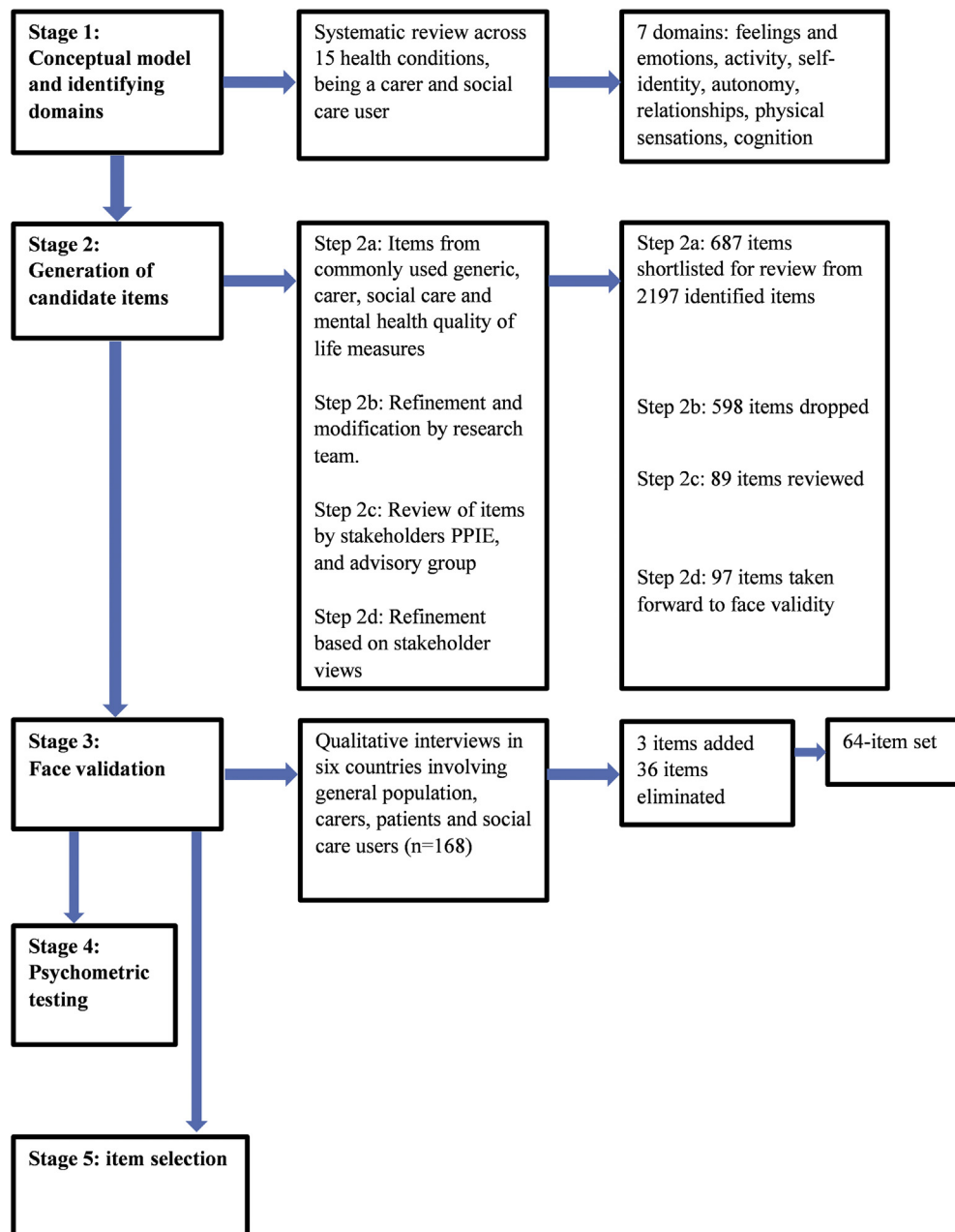
Stage 3: Face Validation

Data collection

Face validation studies were conducted in 6 countries, Argentina, Australia, China, Germany, United Kingdom (UK), and United States of America (US). Semistructured one-to-one cognitive interviews were undertaken with members of the public and carers, patients, and social care users.¹⁴ Participants were asked how they would interpret each question, their ability to respond to it, and their preferences over similar questions with different framing or wording. They were also asked for alternative wording if they highlighted problems with the proposed wording. Each participant saw only a subset of the domains with an overall total of 30 to 50 items. Items were shown in a questionnaire format (Fig. 2). In some cases, different response options could apply, that is, frequency, severity, difficulty, or agree-disagree, and respondents were asked whether they had a preference. All interviewees were provided with training documents and videos and a topic guide (see [Supplemental Materials](https://doi.org/10.1016/j.jval.2021.12.007) found at <https://doi.org/10.1016/j.jval.2021.12.007>). Primary investigators in each country were responsible for ensuring that interviews were undertaken in line with the protocol to ensure a level of consistency internationally. Interviews were conducted in the native language of the participant. A detailed summary of the findings was shared to the wider research team in English. A written informed consent was taken at the start of each interview. Participants completed a short survey (age, sex, ethnicity, any health condition they have, any caring role they have, and 5-level version of EQ-5D), although these questions were not compulsory. At the end of the interview, participants were compensated. All interviews were audio recorded using an encrypted device, and researchers also made brief notes. Ethical approval was obtained from the institutional review boards and relevant ethics committees.

A total of 3 countries (Argentina, China, and Germany) needed translation from English to the respective languages before face validity work. A single translation company undertook the translation following best practice guidelines with forward and back translation by different translators followed by input from the

Figure 1. Overview of the development of the EQ-HWB™.



EQ-HWB indicates EQ Health and Wellbeing; PPIE, patient and public involvement and engagement.

country research team alongside support from the UK team to ensure that the appropriate translations were used (ie, steps 1-6 and 9-10 of the current best practice guidance¹⁵).¹⁶ Topic guides were translated by the country teams.

Participant sample

Patients, social care users, carers (both formal and informal), and members of the general population were invited through different channels in every country (Table 1¹⁷⁻¹⁹).

Data analysis

Data generated from the interviews were analyzed systematically by considering and documenting all feedback/comments

reported by the respondents. Data were recorded on a piloted extraction sheet (see Appendix Table 1 in Supplemental Materials found at <https://doi.org/10.1016/j.jval.2021.12.007>) where item meaning, comprehensibility, item preference, response option preference, and suggested alternatives were recorded. Although interviews were not transcribed verbatim, analysis involved listening to interview recordings and revising notes to ensure immersion in the qualitative data. The researcher that conducted the interview made notes for each item related to the meaning/interpretation of the item, any positive or negative points raised, any suggested alternatives, and preferred items/response options where this was applicable. This information was combined to provide information on the items, including which items to drop

Figure 2. Example of display of items with potential response options.

For each of the following statements, please tick one box that best describes your thoughts, feelings and activities over the last 7 days						
	None of the time	Only occasionally	Some of the time / Sometimes	Often	Most or all of the time	
1 I found it hard to concentrate	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2 I found it hard to focus my thoughts	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3 I found it hard to pay attention	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4 I had trouble thinking clearly	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5 I had trouble remembering	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6 I had trouble with my memory	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7 I felt/was confused	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Alternative Response Sets						
Frequency:	None of the time, Only occasionally, Some of the time / Sometimes, Often, Most or all of the time					
Severity:	Not at all, A little bit, Somewhat, Quite a bit, Very much					
Level of Difficulty:	No difficulty, Slight difficulty, Some difficulty, A lot of difficulty, Unable					
Agreement:	Strongly agree, Agree, Neither agree or disagree, Disagree, Strongly disagree					

(and therefore not be tested in stage 4) and take forward (with or without refinement) to stage 4, and suitability of response options. Each country independently rated each of the items and provided recommendations about which items to retain. The results were then summarized across countries. Self-reported characteristics were used to assess whether particular issues with items arose more in certain groups than others.

Results

Stage 2: Generation of Candidate Items

Step 2a: Mapping of items to domains/subdomains

After reviewing a large pool of items (N = 2197) against the selection criteria, a total of 687 items were collated. Of these, 458 items were extracted from the generic preference and nonpreference-based measures in health and social care and wellbeing measures whereas 229 were drawn from item banks and other measures (see [Appendix Table 2](https://doi.org/10.1016/j.jval.2021.12.007) in Supplemental Materials found at <https://doi.org/10.1016/j.jval.2021.12.007>). Some concepts such as “support,” “stigma,” and “cognition” were identified as being inadequately covered at this stage. Targeted measures and a recent study reviewing measures for assessing wellbeing, happiness, and QoL were used to help identify more items to address these gaps.²⁰

Step 2b: Refinement and modification of items

A more detailed review of the items by the team against the selection criteria resulted in many of the items (n = 598) being dropped from further consideration. There were a number of reasons for dropping items. Many of the items were similar in nature covering the same concepts; for example, different ways of asking about pain and those that were considered to be suitable for a measure that would be used in valuation were selected. There were also items that asked about 2 aspects, for example, impact of pain on functioning that we sought to avoid. In the initial draft item selection, both positively and negatively phrased items were included with further consideration on this issue undertaken in later stages of the project. There was overlap between items related to different subdomains within and across domains. Social engagement items were related to items in other relationship and activity items; autonomy items were related to control and activity items; thinking clearly was related to other cognition items; therefore, these subdomains were not explicitly taken

forward. Items identified for the self-worth/respect subdomain were split into confidence and self-worth subdomains.

Several aspects were taken into consideration around the choice of response options. This included whether or not frequency or intensity best distinguished the level of attainment for a subdomain and the specific wording used. The number of levels was considered based on existing measures, evidence from the literature, and judgment within the research team; a default position of 5 levels was adopted.

Recall periods adopted for self-reported measures vary from today (or yesterday) to last month. The recall period can affect applicability, which may cause missing items (resulting in missing data).²¹ Very short recall periods such as today/yesterday may mean that respondents are not experiencing the issues raised on the particular day.^{12,22} Additionally, capturing broader QoL domains such as coping, control, and loneliness may require a slightly longer recall period. As noted by Norquist et al,²¹ “Longer recall periods may be necessary...when consideration, and integration of events over some period of time is required to reasonably report on the underlying patient reported outcome (PRO) concept (e.g., social functioning).” In contrast, respondents may not remember information accurately over a long recall period and will only report the most salient information rather than “on average.”¹² The need to generate a measure that could be used to track progress after acute events (such as stroke or fracture) in which QoL may change fairly rapidly also makes longer periods of time problematic. A default position of 7 days was adopted at the outset, with regular consideration as to whether this would be most suitable for each item.

Step 2c and d: Review and refinement of items

The results from the face-to-face focus group sessions with the NICE Citizens Council (n = 5) and the PPIE group (n = 7) were combined with the responses from the online survey of advisory group members (n = 28 responses received). Feedback from the consultation frequently focused on adherence/consistency of application of the selection criteria although there was feedback on specific items. Participants provided views on the different items including interpretation and value of including the questions. The advisory group noted “I felt” as more subjective than “I was,” which may also be considered for some items as a clinical diagnosis. The item “I felt cross” was considered problematic by the PPIE and 11 members of the advisory group and hence dropped. Items from the Adult Social

Table 1. Face validation participant demographics.

Country	General public	Carers	Patients	Social care users	Total	Age, range mean (SD)	Female (%)	EQ-5D country tariff utility value	EQ-VAS, mean (SD)
Australia	4	4	17	0	25	28-70 53.7 (14.1)	56	0.848 (0.131)*	N/C
Participants were recruited through an external recruitment company (Stable Research). Purposive sampling was used to include individuals with various physical and mental health conditions, carers, and members of the general public.									
Argentina	8	8	0	8	24	24-91 54 (20)	63	N/C	N/C
Participants were recruited using different strategies. Known individuals were contacted (through local researchers' informal networks). A snowball sampling approach was adopted asking participants to help researchers to identify further individuals and particularly social care users. Finally, we visited health promotion public facilities in the city of Buenos Aires ("Estaciones Saludables") to recruit users of those services.									
China	0	13	17	0	30	18-71 37.73 (15.55)	60	N/C	N/C
Participants were recruited using a convenience sampling approach from 2 hospitals in Shanghai, No.10 Hospital of Shanghai and Zhongshan Hospital of Fudan University. Most participants were recruited from the outpatient services; some were recruited from inpatient services.									
England	6	13	18	8	45	23-95 60.4 (20.2)	58	0.78 (0.23)†	N/C
Participants with physical health conditions were recruited from Sheffield Teaching Hospital Patient panels (Cardiovascular Patient Panel, Diabetes and Endocrinology Panel, Therapeutics and Palliative Care Panel, Online Public Advisory Panel, Motor Neuron Disease Panel, Stroke Panel). Mental health service users were recruited through RDaSH targeting mental health service users including those receiving drug and alcohol rehabilitation. Social care users were recruited through a day center and residential care home (via Doncaster City Council). Carers were recruited through Sheffield Carers Centre via an email to their list and an advert on their website. Members of the general public were recruited through the University of Sheffield volunteers list for staff but excluding academic staff and the School of Health and Related Research (where the research was conducted).									
Germany	0	12	8	7	27	21-30 yrs n = 6 31-40 yrs n = 6 41-50 yrs n = 4 51-60 yrs n = 7 61-70 yrs n = 2 71-80 yrs n = 2	70	0.85 (0.20)‡	73.50 (19.68)
Participants were recruited in 2 hospitals, a rehabilitation clinic and a physiotherapy practice in Bielefeld and Berlin, and at Bielefeld University. A purposive sampling approach was used to include 3 key groups of interest: patients (mental and physical conditions), social care users, and carers (formal and informal).									
US	0	0	19	0	19	23-76 53.8 (13.8)	53	0.84 (0.20)‡	77.3 (14.78)
Respondents with acute and long-term physical and mental health conditions were recruited from clinics at the University of Illinois Hospital and Health Sciences System and the website ResearchMatch.org .									

N/C indicates not collected; No., number; RDaSH, Rotherham Doncaster and South Humber; US, United States of America; VAS, visual analog scale.

*Based on Devlin et al.¹⁷

†Based on Ludwig et al.¹⁸

‡Based on Pickard et al.¹⁹

Care Outcomes Toolkit were identified as problematic for generic use because they were tailored toward recipients of care. The subdomains around guilt/shame and burden were dropped during early consultation because of social desirability concerns. Further details of the results of the PPIE results are presented in the Supplemental Materials found at <https://doi.org/10.1016/j.jval.2021.12.007>. A total of 97 items were taken forward into face validation.

Stage 3: Face Validation

Face validation studies were conducted between April 2018 and February 2019. Participant characteristics for the face validity study for each of the participating countries are presented in Table 1.¹⁷⁻¹⁹ A total of 170 interviews were conducted with

patients (n = 79), social care users (n = 23), carers (both formal and informal, n = 50), and members of the general population (n = 18).

A summary of the common and core findings for each of the 7 domains is outlined below and summarized in Table 2.

Domain Specific-Findings

Of the 97 draft items taken into stage 3, 36 items were eliminated based on the evidence in this stage. A total of 3 additional items were added. This resulted in a 64-item set (Fig. 3).

Activity

This domain aimed to capture functioning and covered self-care, enjoyable or meaningful activities/roles, mobility, communication (speech), hearing, and vision. A total of 24 potential items

Table 2. Results of face validation studies.

Item	UK	Argentina	Australia	China	Germany	US	Outcome (K/M/D)	Item taken forward
Domain: activity								
I enjoyed what I did (F).				∅			K	
I was able to do the things I value (F).				∅			K	
I did things I found rewarding.				∅			D	
I was bored.							D	
I did what I wanted to do.				∅	∅		D	
I could do the things I wanted to do (F).				∅			K	
I did what I needed to do.							D	
I was able to do what I needed (F).				∅	∅		K	
I had no difficulty with my day-to-day activities/daily activities.		∅		∅	∅	∅	M	How well were you able to do your day-to-day activities (eg, working, shopping, traveling) (D)?
Given the help I had/received, my personal needs were met (eg, being washed, going to the toilet, getting dressed, having food when I needed).				∅	∅		M	My personal needs were met (eg, being washed, going to the toilet, getting dressed, having food when I needed) (F).
Given the help I had/received, my self-care needs were met (eg, being washed, going to the toilet, getting dressed, having food when I needed).		∅		∅	∅		D	
I was able to look after myself (F).	∅			∅	∅		K	
I needed help with looking after myself.				∅	∅		D	
I was able to look after myself with no difficulty.	∅	∅			∅	∅	M	I was able to look after myself (eg, being washed, going to the toilet, getting dressed, having food when I needed) (F).
I had no difficulty with self-care activities.		∅			∅	∅	D	
I was able to get around inside my home with no difficulty (D).						∅	K	
I was able to get around outside with no difficulty (D).	∅				∅	∅	K	
How well did you communicate with others?		∅	∅		∅		D	
I was able to communicate with others with no difficulty.		∅				∅	D	
Because of hearing and/or speech, how difficult did you find it to have a conversation (D)?		∅		∅	∅		K	
How well can you hear (using hearing aids if needed)?				∅	∅		M	How well can you hear (using hearing aids if you usually wear them) (D)?
I had no difficulty hearing (using hearing aids if needed).				∅		∅	D	
How well can you see (using your glasses or contact lenses if they are needed) (D)?				∅	∅		K	

continued on next page

Table 2. Continued

Item	UK	Argentina	Australia	China	Germany	US	Outcome (K/M/D)	Item taken forward
I had no difficulty seeing (using your glasses or contact lenses if they are needed).				∅		∅	D	
								New item: I was able to do the things I wanted to do (S).
Domain: autonomy								
I felt able to cope.					∅	∅	M	I felt able to cope with my day-to-day life (F).
I felt unable to cope.							D	
I felt unable to cope with my day-to-day life (F).				∅	∅	∅	K	
I felt overwhelmed by my problems.				∅	∅		M	I felt overwhelmed by the problems or situation (F).
I felt in control of my daily life.		∅		∅	∅		D	
I felt in control of my day-to-day life (F).				∅			K	
I have as much control over my daily life as I want.	∅	∅		∅	∅		M	I had control over my day-to-day life (F).
								New item: I felt I had no control over my day-to-day life (F).
Domain: cognition								
I found it hard to concentrate (F).			∅	∅			K	
I found it hard to focus my thoughts.					∅	∅	D	
I found it hard to pay attention (F).							K	
I had trouble thinking clearly (F).					∅		K	
I had trouble remembering (F).			∅				K	
I had trouble with my memory.				∅	∅		D	
I felt confused (F).	∅	∅	∅	∅			K	
Domain: feelings and emotions								
I felt happy (F).				∅			K	
I felt unhappy (F).		∅		∅			K	
I felt depressed.				∅	∅		D	
I felt sad (F).	∅			∅			K	
I enjoyed life.		∅		∅			D	
I felt content with my life.							D	
I thought my life was not worth living (F).				∅			K	
I felt that I had nothing to look forward to (F).	∅				∅		K	
I had nothing to look forward to.				∅	∅		D	
I looked forward to each day.		∅		∅			D	
I felt frightened (F).	∅						K	
I felt afraid (F).	∅						K	
I felt safe (F).		∅		∅	∅		K	
I felt unsafe (F).				∅	∅		K	
I felt secure.				∅	∅		D	
I felt anxious (F).		∅		∅	∅		K	

continued on next page

Table 2. Continued

Item	UK	Argentina	Australia	China	Germany	US	Outcome (K/M/D)	Item taken forward
My worries overwhelmed me.					∅		D	
I felt worried (F).				∅			K	
I felt calm (F).							K	
I felt relaxed.							D	
I felt irritable (F).		∅					K	
I felt irritated.		∅					D	
I felt angry (F).							K	
I felt frustrated (F).							K	
I lost my temper easily (F).					∅		K	
								New item: I felt cheerful (F).
Domain: physical sensations								
I had no pain (mild pain, etc).				∅		∅	M	I had no physical pain (mild pain, etc) (S).
How often do you experience pain?	Ü			∅		∅	M	How often do you experience physical pain (F)?
I had no discomfort (mild discomfort, etc).				∅	∅	∅	M	I had no physical discomfort (mild discomfort, etc) (S).
How often do you experience discomfort?				∅	∅	∅	M	How often do you experience physical discomfort (F)?
I felt exhausted (F).			∅				K	
I got tired easily.							M	I felt very tired (F).
I was too tired to do anything.							D	
I had problems with my sleep (F).							K	
Domain: relationships								
I felt supported by other people.		∅	∅	∅			D	
I felt unsupported (F).			∅		∅	∅	M	I felt unsupported by people (F).
Other people gave me support.		∅		∅			D	
I had support when I needed it (F).							K	
I had disagreements and conflict with people.				∅			D	
I got on with people around me.		∅	∅		∅	∅	M	I got along well with people around me (F).
I got along well with people I came into contact with.		∅			∅	∅	D	
I felt lonely (F).							K	
I felt there was nobody I was close to (F).	∅						K	
I felt I had no one to talk to (F).				∅	∅		K	
I felt isolated (F).	∅	∅			∅		K	
I felt people avoided me (F).	∅						K	
I felt judged by others.					∅		D	
I felt accepted by others (F).							K	
I felt excluded (F).						∅	K	
I felt left out (F).							K	
Domain: self-identity								
I felt confident in myself (F).				∅	∅		K	

continued on next page

Table 2. Continued

Item	UK	Argentina	Australia	China	Germany	US	Outcome (K/M/D)	Item taken forward
I felt confident.		∅		∅			D	
I felt unsure about myself (F).							K	
I felt I was treated with respect.							D	
I felt respected.					∅		D	
I felt like I lived with dignity.				∅			D	
I felt good about myself (F).				∅	∅		K	
I felt like a failure (F).	∅						K	
I felt valued.					∅		D	
I felt useful.				∅	∅		D	

∅ indicates mixed evidence; ∩, problems identified; ∪, no problems identified; D, drop; F, frequency response option; K, keep; M, modify; S, severity response option, UK, United Kingdom; US, United States of America.

were tested and 11 were dropped, whereas 1 was added (Fig. 3). Questions that referred to what individuals “wanted” to do versus “needed” were interpreted correctly with the former referring to what was preferred and the latter to activities that were essential such as activities of daily living. Nevertheless, for some items, there was ambiguity because of differences in interpretation, brevity, and the lack of context. For example, some items were interpreted in different ways to what was intended, for example, “communicate” inferred to mean methods of communication—telephone, conversation, text, and email; skill in getting a message across effectively; and the response of others (eg, clinical staff not listening to them). This does not link to the original construct of hearing and speaking and points to ambiguity as to what respondents’ answers would be referring to. The term “self-care” was not commonly used to mean things such as washing/dressing. In mental health, self-care was interpreted to mean the things that they did to improve their wellbeing, rather than in terms of physical self-care (ie, washing, dressing). Similarly, self-care was seen as arising from both physical limitations and resource limitations (eg, lack of time).

Including aspects of “receiving help” was problematic even in groups where help could have been received (ie, patients); therefore, this was rephrased. The items aimed to distinguish between personal care outcomes attained over the last week (what actually happened) and the respondent’s ability to attain personal care outcomes independently (what they would have achieved if they did not have care/support). These items created ambiguity in interpretation from respondents who did not receive any care or would have benefited from additional care/support.

The relevance of some items was also highlighted. This included comments around what could be reasonably expected, for example, “everyone experiences boredom” or “unrealistic to expect people to be able to do what they want.” There were also issues with questions related to self-care and receiving help for some carers who did not know why they would be asked these questions.

Autonomy

This domain covered coping and control and was mainly testing different ways of asking the same question. A total of 7 potential items were tested, 2 were dropped, and 1 new item was added (Fig. 3). There was a preference for items that had more information, for example, coping with day-to-day life rather than just coping. An item that provided a definition of control was found to be helpful by many of the respondents.

Cognition

Concentration, memory, and confusion were covered in this domain. A total of 7 potential items were tested and 2 were dropped (Fig. 3). Most participants understood the questions and said they would be able to answer them. “Memory” was considered to be a long-term issue and not something in the context of 7 days. Some respondents interpreted this to be referring to dementia with some questioning whether this would be something that could be answered, that is, “would I know that I have memory loss.”

Feelings and emotions

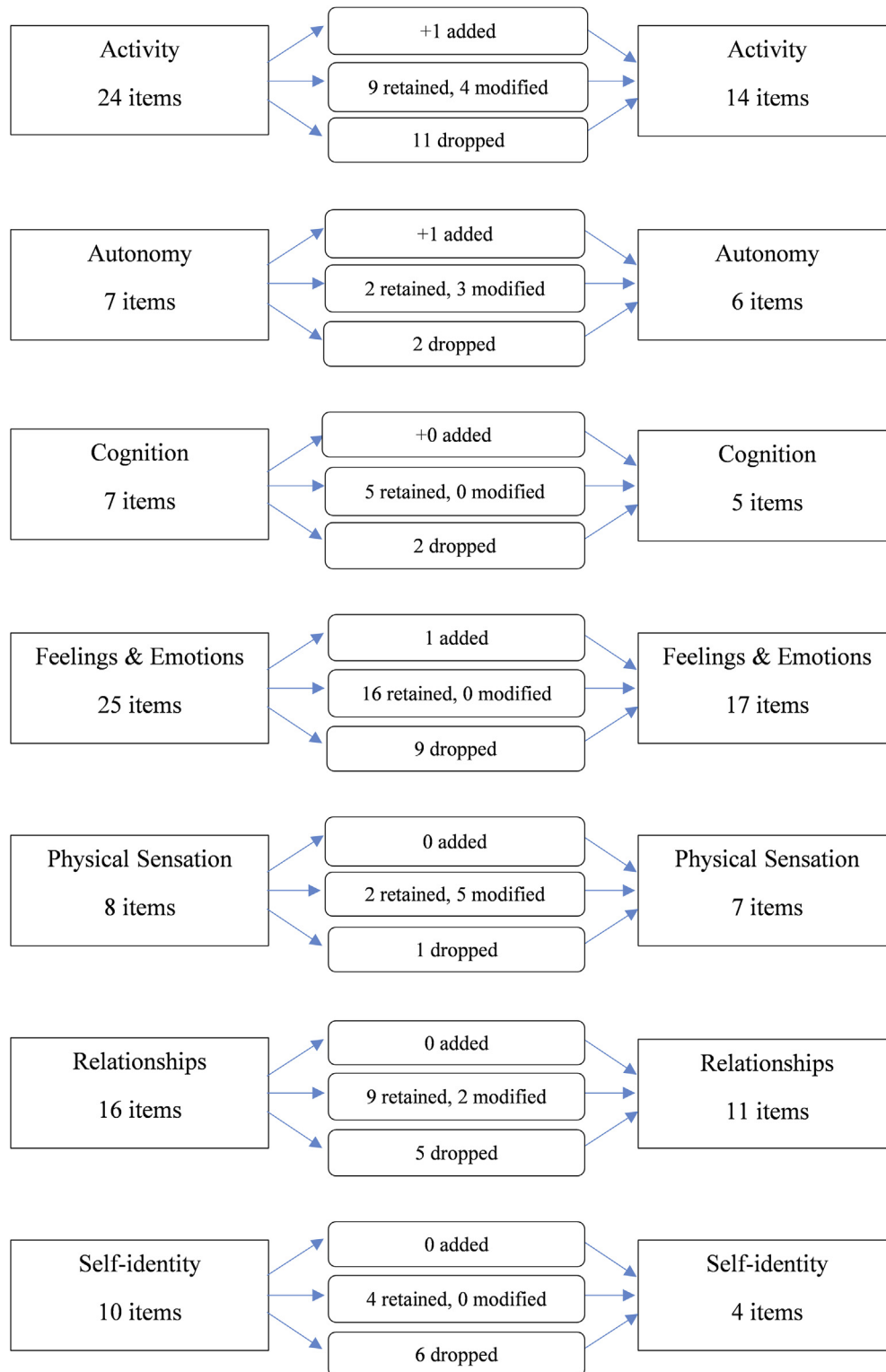
This domain covered sadness, happiness, worry, hope and hopelessness, anger and frustration, vulnerability and safety, and guilt/shame. A total of 25 potential items were tested and 9 were dropped (Fig. 3). Many of the items were interpreted correctly and respondents could answer them, although there were issues with some. In the happiness/depression subdomain, some respondents felt that the top end of “happy” and “enjoyed life” was unrealistic; that is, “no one enjoys life all the time.” The term “depressed” was interpreted to mean having a clinical diagnosis by some respondents. In the hope/hopelessness subdomain, the item on “life not worth living” was considered quite negative. Looking forward to each day was not considered to be something that individuals did every single day, whereas “look forward to” needed further information in some countries. “Safe” and “secure” were considered to be ambiguous terms in the safety subdomain whereas “relaxed” was considered to be a physical state in the anxiety/calm domain.

Physical sensations

This domain covered pain, discomfort, sleep problems, and fatigue. A total of 8 potential items were tested, most of which performed well in face validity and only 1 was dropped (Fig. 3). Discomfort was often interpreted to include mild pain. The term “physical” was added to pain and discomfort items to distinguish this from mental health-related aspects.

Relationships

This domain covered loneliness, social engagement, stigma, support, positive relationships and relationships, belonging and connectedness, and burden to others. A total of 16 potential items were tested, with many performing well in terms of interpretation and ability to respond to them and only 5 were dropped (Fig. 3).

Figure 3. Summary of item modification after face validation.

Social support framed as “support” or “by other people” resulted in some ambiguity. “Support” was unclear whereas “other” resulted in respondents considering people who were not those they saw regularly. “Disagreements and conflict” was considered

problematic because it focused on 2 issues and had mixed interpretation in terms of impact on QoL given that some respondents thought of it as positive to be able to have disagreements (UK and Australia). “Got on” was colloquial and did not translate well. The

term “judged” was also ambiguous and not necessarily negative in all interpretations.

Self-identity

This domain aimed to cover feelings of confidence and self-worth and being treated with dignity/respect. A total of 10 items were tested and 6 were dropped including 1 subdomain, dignity/respect (Fig. 3). “Confidence” had broad interpretations some of which were relevant. Nevertheless, many of the other items in this domain were problematic. Dignity was linked to respondents’ own behavior rather than the behavior of others whereas respect was linked to manners or very specific incidents. Therefore, this subdomain was dropped. “Feeling valued/useful” was not relevant to older people because of how the terms were interpreted, that is, doing tasks or being paid. “Feeling good” had some irrelevant interpretations, for example, “how I look,” whereas others were related to physical health, that is, “I felt well.”

Common Findings

Respondents found it useful to have examples of the construct being measured—and this was a common finding across the different domains. Brief items could be answered but respondents wanted information on context and this was true across different countries. There was also a preference for simpler layouts in presenting questions.

Although there were some differences in response option preferences, for example, frequency over severity, this was often mixed and respondents were often unable to say why they preferred one option to another. Recall periods were sometimes considered too short for particular constructs such as coping and control or irrelevant such as hearing where the loss is permanent. Completion instructions for the draft measure, including the recall period, were usually displayed at the top of the page or table. These were often ignored or forgotten by participants.

Combining the Evidence to Inform the Content of the Psychometric Survey (Stage 4)

The results of stages 2 and 3 were used to inform the selection of items taken forward to stage 4 (psychometric survey)⁹ (Table 2).

Discussion

This project aimed to develop a broader generic measure of QoL for use in economic evaluation that would be relevant for use across health and social care. Methods of development drew upon current good practice for measure development, covering multi-country, multilingual, and multicultural considerations.^{4,13,15,23} The generation of items based on terms from the qualitative review⁸ and items from existing health and wellbeing measures resulted in 687 candidate pool of items from a list of 2197 potential items. Items were identified for 28 subdomains across 7 domains. This approach allowed for full consideration of the relevance, comprehensiveness, and comprehensibility (ie, content validity) of the new measure.

Stage 3 incorporated an ambitious multicountry face validation exercise to further test and examine the suitability of the proposed item pool and response options. A total of 97 items were tested in the face validation and 47 items were retained, and 14 were modified whereas 3 were added to the candidate pool of items for consideration in further stages. One subdomain was dropped. The approach benefited early in the development phase of the measure from a multiculture, multilingual approach with common

methodology used across different countries, which was important in considering wider audiences who may use the measure.

The results were used to help inform the reduction of the item pool to take forward to stage 4 (psychometric survey) and were used as evidence to inform final item selection for the EQ-HWB measure. Many items were identified as being potentially problematic in face validity interviews across the different groups. Short items without additional context raised concerns and uncertainties about their scope yet longer items risked problems with readability. Using different population groups was important because some items worked better in some groups than others. For example, being able to communicate well, from a patient perspective, has a physical emphasis; for some nonpatients/carers, this is interpreted as how successfully they reveal communication skills.

The project was not without its challenges. Logistical difficulties associated with ethical and governance approval processes across the included countries made iterative decision making challenging. Although general population, patient, carer, and social care perspectives were sought across the whole project, this was not achieved for all countries. Recruitment from social care was completed in 3 of the 6 countries (Argentina, England, and Germany). The steps undertaken in the development of potential items and response options were robust and followed recognized best practice. This study did not undertake a qualitative study to generate items as advocated in the Consensus-Based Standards for the Selection of Health Measurement Instruments.²⁴ Instead, data from existing evidence (including published qualitative reviews and established measures of health and wellbeing) were used, which had the advantage of drawing from a broader range of voices including different mental and physical health patient groups, carers of different types of individuals, and users of social care. Audio recordings from discussions with PPIE or stakeholder groups and face validation studies were not transcribed verbatim as recommended in the Consensus-Based Standards for the Selection of Health Measurement Instruments.²⁴ Although verbatim transcription was not undertaken, audio recordings were used to complete data extraction from the interviews themselves. Given the tight focus of the interviews on cognitive debriefing of predetermined items, transcription was not considered necessary. Resource and time implications were considered; nevertheless, the primary reason was one of minimizing research waste and the ethical implications of undertaking research with no clear rationale. It was viewed to be more important to check interpretation across a broad sample.

Conclusions

A candidate pool of items was identified and selected for testing in face validation across 6 countries to cover a broad range of content important to patients, social care users, and informal carers worldwide. In these initial stages, we exhaustively searched items, mapped them to domains and subdomains, and carried forward a successful face validation of an initial item pool. Although there were some discrepancies among 6 countries, there were useful common findings to select items for the next stage. In doing this, items were identified that were considered appropriate and understandable across all included groups of participants and across different countries and cultural contexts. The international evidence was used to support decision making for item retention and elimination for subsequent stages of the EQ-HWB development. The EQ-HWB has a potential for becoming a valuable addition to the supply of QoL measures in research and economic evaluation across health, social care, and public health worldwide.

Supplemental Material

Supplementary data associated with this article can be found in the online version at <https://doi.org/10.1016/j.jval.2021.12.007>.

Article and Author Information

Accepted for Publication: December 8, 2021

Published Online: February 25, 2022

doi: <https://doi.org/10.1016/j.jval.2021.12.007>

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

Correspondence: Jill Carlton, PhD, School of Health and Related Research, University of Sheffield, Regent Crt, 30 Regent St, Sheffield, England S1 4DA, United Kingdom. Email: j.carlton@sheffield.ac.uk

Author Contributions: *Concept and design:* Carlton, Peasgood, Mukuria, Connell, Brazier, Engel, Pickard

Acquisition of data: Carlton, Peasgood, Connell, Ludwig, Marten, Kreimeier, Engel, Belizán, Yang, Monteiro, Kuharic, Luo, Mulhern, Greiner, Pickard, Augustovski

Analysis and interpretation of data: Carlton, Peasgood, Mukuria, Connell, Brazier, Ludwig, Marten, Kreimeier, Engel, Belizán, Yang, Monteiro, Kuharic, Mulhern, Greiner, Pickard, Augustovski

Drafting of the manuscript: Carlton, Mukuria, Brazier, Engel, Mulhern

Critical revision of the paper for important intellectual content: Carlton, Peasgood, Mukuria, Brazier, Ludwig, Marten, Kreimeier, Engel, Belizán, Yang, Monteiro, Kuharic, Luo, Mulhern, Greiner, Pickard, Augustovski

Statistical analysis: Monteiro, Kuharic, Augustovski

Provision of study materials or patients: Carlton, Ludwig, Marten, Kreimeier, Kuharic, Mulhern, Greiner, Augustovski

Obtaining funding: Carlton, Peasgood, Mukuria, Brazier, Ludwig, Marten, Kreimeier, Engel, Yang, Monteiro, Luo, Mulhern, Greiner, Pickard

Administrative, technical, or logistic support: Carlton, Ludwig, Marten, Kreimeier, Monteiro, Kuharic, Greiner

Supervision: Brazier, Pickard

Conflict of Interest Disclosures: Drs Carlton, Peasgood, Mukuria, Brazier, Ludwig, Marten, Kreimeier, Engel, Belizán, Yang, Luo, Grenier, and Mulhern and Mses Connell, Monteiro, and Kuharic reported receiving grants from the EuroQol Research Foundation during the conduct of the study. Drs Peasgood, Mukuria, Brazier, Ludwig, Marten, Kreimeier, Engel, Yang, Luo, Mulhern, and Grenier reported being members of the EuroQol Research Association. Drs Mukuria and Brazier reported receiving grants from the EuroQol Research Foundation outside the submitted work. Dr Brazier reported receiving grants from the UK Medical Research Council during the conduct of the study and reported receiving grants and personal fees and previously being a member of the EQ Group Executive outside the submitted work and having a patent for SF-6D and SF-6Dv2 with royalties paid to the University of Sheffield. Drs Ludwig, Marten, Kreimeier, Engel, and Grenier reported receiving nonfinancial support from the EuroQol Research Foundation during the conduct of the study. Miss Kuharic reported receiving a fellowship from Takeda Pharmaceuticals fellowship for her graduate studies outside the submitted work. Dr Luo reported receiving personal fees from the EuroQol Research Foundation outside the submitted work. Drs Mulhern and Luo are editors for *Value in Health* and had no role in the peer-review process of this article. Drs Pickard and

Augustovski reported receiving a grant payment to the institution (university) from the EuroQol Research Foundation. No other disclosures were reported. The views expressed in this manuscript are those of the authors and not necessarily of our funders, the National Institute for Health and Care Excellence, the Department of Health and Social Care, or those acknowledged.

Funding/Support: This study is independent research funded by the UK Medical Research Council (grant number 170620) and 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 thank the National Institute for Health and Care Excellence for highlighting the methodological research need to the Medical Research Council that resulted in the funding call entitled "Beyond the QALY," which led to this research being funded. The authors acknowledge the support of the National Institute for Health Research Yorkshire and Humber Applied Research Collaboration (formerly CLAHRC) and the National Institute for Health Research Clinical Research Network. We 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 the EuroQol Group Association for their input at plenary and academy meetings and the EuroQol office for their support. Finally, the authors acknowledge the contribution of all the patients, social care users, and informal carers who took part in all the studies across the different country.

REFERENCES

- McDowell I, Newell C. *Measuring Health: a Guide to Rating Scales and Questionnaires*. 2nd ed. New York, NY: Oxford University Press; 1996.
- Connell J, Carlton J, Grundy A, et al. The importance of content and face validity in instrument development: lessons learnt from service users when developing the Recovering Quality of Life measure (ReQoL). *Qual Life Res*. 2018;27(7):1893–1902.
- Johnson RL, Morgan GB. *Survey Scales: A Guide to Development, Analysis, and Reporting*. New York, NY: Guilford Publications; 2016.
- Guidance for industry: patient-reported outcomes measures: use in medical product development to support labeling claims. US Department of Health and Human Services, Food and Drug Administration. <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM19328.pdf>. Accessed May 19, 2021.
- Brazier JE, Peasgood T, Mukuria C, et al. The EQ Health and wellbeing: overview of the development of a measure of health and wellbeing and key results. *Value Health*. In press.
- Mitchell PM, Al-Janabi H, Richardson J, Iezz A, Coast J. The relative impacts of disease on health status and capability wellbeing: a multi-country study. *PLoS One*. 2015;10(12):e0143590.
- Brazier JE, Rowen D, Lloyd A, Karimi M. Future directions in valuing benefits for estimating QALYs: is time up for the EQ-5D? *Value Health*. 2019;22(1):62–68.
- Mukuria C, Connell J, Carlton J, et al. Qualitative Review 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. *Value Health*. In press.
- Peasgood T, Mukuria C, Brazier J, et al. Developing a new generic health and wellbeing measure: psychometric survey results for the EQ Health and Wellbeing. *Value Health*. In press.
- 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(7):1067–1081.
- 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.
- Bradburn NM, Sudman S, Wansink B. *Asking Questions: The Definitive Guide to Questionnaire Design—for Market Research, Political Polls, and Social and Health Questionnaires*. Chichester, United Kingdom: John Wiley & Sons; 2004.
- Streiner DL, Norman GR, Cairney J. *Health Measurement Scales: a Practical Guide to Their Development and Use*. 5th ed. Oxford, United Kingdom: Oxford University Press; 2014.
- Tourangeau R. Cognitive science and survey methods: a cognitive perspective. In: Jabine T, Straf M, Tanur J, Tourangeau R, eds. *Cognitive Aspects of Survey Design: Building a Bridge Between Disciplines*. Washington, DC: National Academy Press; 1984:73–100.

15. Arafat SM, Chowdhury H, Qusar MMA, Hafez MA. Cross cultural adaptation and psychometric validation of research instruments: a methodological review. *J Behav Health*. 2016;5(3):129–136.
16. Linton MJ, Dieppe P, Medina-Lara A. Review of 99 self-report measures for assessing well-being in adults: exploring dimensions of well-being and developments over time. *BMJ Open*. 2016;6(7):e010641.
17. Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: an EQ-5D-5L value set for England. *Health Econ*. 2018;27(1):7–22.
18. Ludwig K, Graf von der Schulenburg JM, Greiner W. German value set for the EQ-5D-5L. *Pharmacoeconomics*. 2018;36(6):663–674.
19. Pickard AS, Law EH, Jiang R, et al. United States valuation of EQ-5D-5L health states using an international protocol. *Value Health*. 2019;22(8):931–941.
20. Morris Devlin, Parkin Spencer. *Economic Analysis in Healthcare*. 2nd ed. Hoboken, NJ: Wiley; 2012.
21. Norquist JM, Girman C, Fehnel S, DeMuro-Mercon C, Santanello N. Choice of recall period for patient-reported outcome (PRO) measures: criteria for consideration. *Qual Life Res*. 2012;21(6):1013–1020.
22. Wild D, Grove A, Martin M, et al. Principles of good practice for the translation and cultural adaptation process for Patient-Reported Outcomes (PRO) measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value Health*. 2005;8(2):94–104.
23. Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2—assessing respondent understanding. *Value Health*. 2011;14(8):978–988.
24. Gagnier JJ, Lai J, Mokkink LB, Terwee CB. COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. *Qual Life Res*. 2021;30(8):2197–2218.