

Quantitative testing of potential modifications to item wording and presentation in the EQ-HWB long form

Stevanus Pangestu^{ID} · Cate Bailey · Clara Mukuria · Tessa Peasgood · Alex Bató · Eszter Szlávicz · Fanni Rencz

Received: 10 June 2026 / Accepted: 19 August 2026

© The Author(s) 2026

Abstract

Objectives This study compared the psychometric performance of alternative item wording and formats in the experimental EQ-HWB long form: (1) presentation formats for positively worded items, (2) positively worded items from EQ-HWB v1.1 versus negatively worded counterparts from EQ-HWB v1.2, (3) hopelessness item wording, and (4) vision item wording.

Methods Overall, 1,200 general-population adults completed the experimental modified EQ-HWB (v1.2) item-by-item with additional item variants and were randomized to three presentation formats for positively worded items: text warning, reversed response scale, or unmodified. Positively worded items ('accepted', 'feeling good', 'do things one wanted to') were compared with negatively worded counterparts ('excluded', 'negative feelings', 'difficulties with enjoyable activities'). To identify response errors from negative-to-positive wording switches, positive-item responses were benchmarked against external measures (e.g., happiness). 'Hopelessness' was compared with 'nothing to look forward to', and vision items with and without reference to aids. Psychometric performance was assessed using distributional properties, and convergent and known-group validity analysis.

Results Response errors for positively worded items occurred across formats but were higher in the unmodified format (5.8–6.8%) than across alternative formats (3.3–4.5%). These items showed lower ceilings versus negatively worded counterparts and slightly improved instrument-level known-group validity. 'Hopelessness' showed stronger known-group validity than 'nothing to look forward to'. Referencing aids slightly reduced reported difficulties, while both vision versions similarly distinguished those with chronic eye disease.

Conclusions Our findings can inform key design decisions on item wording and presentation in the EQ-HWB long form and should be considered alongside qualitative evidence from the literature.

Keywords EQ-HWB · Health · Long form · Psychometrics · Item framing · Well-being

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.

Précis: This study compares the psychometric performance of alternative item wording and presentation formats in the long form of the experimental EQ-HWB instrument.



Introduction

The EuroQol Health and Wellbeing (EQ-HWB) is an experimental, generic measure developed for populations in which assessing well-being alongside health is relevant, such as capturing the impacts of receiving or providing care. The instrument was developed by international researchers from the EuroQol Group in collaboration with the University of Sheffield as part of the ‘Extending the QALY’ (E-QALY) project [1, 2]. Two versions have been developed: the EQ-HWB-9, a shorter classification system intended for use in cross-sectoral economic evaluation and decision-making, and the EQ-HWB long form, a profile measure that extends the EQ-HWB-9 with a broader set of items. While the EQ-HWB-9 is approaching finalization, the long form remains under development, with ongoing international qualitative and quantitative testing of modifications.

The initial 25-item experimental version of the EQ-HWB long form (hereafter the ‘experimental’ version, v1.1) was released in 2022 [1]. Evidence from a series of qualitative and quantitative studies informed the updated 25-item version released in 2024 (hereafter the ‘experimental modified’ version, v1.2) [3–19]. Compared with the original experimental version, the modified version introduced changes to the instructions, item ordering, response levels, item stems, and wording. To date, quantitative evidence on the experimental modified EQ-HWB(-9) remains limited to one study each in Australia and China [20, 21], while qualitative evidence comes from Australia and Hungary [7, 22, 23].

An important consideration for the finalization of the long form concerns the three items for which alternative positive and negative wordings have been considered. One content validity study of the EQ-HWB in Australia reported mixed views on the positively worded items: some respondents preferred their use, particularly when placed at the end of the questionnaire to leave a more positive overall impression; while others reported difficulties in interpreting the response levels, which were identical to those used in the earlier negatively worded questions [22]. Given the limited evidence, it remains unclear whether differences between positively and negatively worded items reflect respondents overlooking response options, difficulties in comprehension, differences in the constructs being measured, or other factors. If positive items are to be retained, mitigation strategies may be needed to reduce unintended response errors.

In the EQ-HWB-9, the ‘getting around inside or outside’ item was finalized with wording that refers to the use of aids. Referencing aids may help clarify the intended meaning and support more accurate responses; however, previous qualitative evidence indicated that some respondents who did not use aids misunderstood the item, interpreting it as asking whether they used aids [7, 22]. Similar issues may arise for the vision item in the long form, which currently includes such a reference (e.g., “how much difficulty did you have seeing? *Using, for example, glasses or contact lenses if you normally use them*”). Another change in the EQ-HWB long form requiring testing is the replacement of ‘nothing to look forward to’ with ‘hopelessness’, prompted by early translation and cognitive debriefing evidence from the EuroQol Version Management Committee.

Therefore, this study aimed to compare the psychometric performance of alternative item wording and presentation formats in EQ-HWB long form across four areas: (1) different presentation formats for positively worded items to mitigate potential response errors; (2) three positively worded items from the experimental version (v1.1) and their negatively worded counterparts from the experimental modified version (v1.2); (3) wording of the ‘hopelessness’ item; and (4) ‘vision’ item versions with and without reference to aids. Findings are expected to inform decisions on item modifications and finalization of the EQ-HWB long form.

Methods

Study design and respondents

This study used data from a cross-sectional online survey of 2,000 adults from the Hungarian general population, conducted in August–September 2025. Respondents were recruited through two online panels using non-probabilistic quota sampling to approximate national representativeness by sex, age group (18–34, 35–44, 45–54, 55–64, 65+ years), education (primary, secondary, tertiary), region (Central, Western or Eastern Hungary), and residence type (capital, other city, village).

Respondents were randomly allocated to five study arms using matching criteria based on age group, sex, education, region, residence type, and general health status (excellent/very good, good, fair/poor). The present study included the three arms that completed the EQ-HWB long form items ($n = 400$ each; total $n = 1,200$). The remaining two arms served different objectives, did not include the EQ-HWB long form, and were not analyzed in this article [24]. The survey was anonymous, and respondents received redeemable reward points as incentives. Ethical approval for the data collection was granted by the Research Ethics Committee of Corvinus University of Budapest (KRH/284/2024).

Survey

The survey was administered in Hungarian, with all questions mandatory and no missing responses. The main outcome measure was the 2024 modified version of the EQ-HWB long form (v1.2), supplemented with additional item variants. Additional self-reported outcomes included the EQ-5D-5 L [25]; general health, general well-being, and general quality of life (each assessed on a 5-point Likert scale); happiness (0–10 numeric rating scale); the Rosenberg Self-Esteem Scale [26]; and the second item of the InCharge Financial Distress/Financial Well-Being Scale, measuring overall perceived financial satisfaction (0–10 ladder scale) [27]. In addition to the quota variables, sociodemographic variables included marital status, employment status, and monthly net household income. Health and lifestyle-related variables included body weight and height, caregiving and care recipient status, physician-diagnosed chronic conditions (selected from a list of 35), and use of vision aids in the last 7 days (frequency and type). To ensure data quality, several checks were implemented, including randomly placed attention-check questions, repeated reporting year of birth, inclusion of two fictitious conditions in the chronic condition list, and response time (speed) checks. Respondents failing to meet any of these criteria were excluded.

EQ-HWB and additional item variants

The EQ-HWB and its item variants were administered in a customized item-by-item format across four blocks, with a fixed item order within each block. Other questionnaire items were interspersed between blocks. The first two blocks comprised the experimental modified version (v1.2) of the EQ-HWB; block 1 included the nine EQ-HWB-9 items, and block 2 included the remaining 16 long-form items. Blocks 3 and 4 included additional item variants. Block 3 comprised nine items drawn from an earlier EQ-HWB version (v1.1), modification testing in other methodological studies, and the original instrument development item pool [2, 21, 28], and one item developed specifically for the present study (i.e., ‘vision’ item without aid reference). Block 4 included three positively worded items from the experimental version (v1.1) of EQ-HWB: ‘feeling accepted’, ‘feeling good about oneself’, and ‘doing things one wanted to do’ [1]. Each item was assessed using a five-level response scale reflecting either difficulty (no, slight, moderate, a lot of difficulty, or unable), frequency (none of the time, a little of the time, sometimes, often, or most or all of the time), or severity (no, mild, moderate, severe, or very severe). At the instrument level, the EQ-HWB long form was scored using a level sum score (LSS), calculated by summing responses across all 25 items. This yielded a theoretical score of 25 to 125, with higher scores indicating

worse EQ-HWB outcomes. A complete list and ordering of all administered EQ-HWB items and their response options used in this study are provided in Appendix Table 1, along with their position in the overall survey order.

Presentation of positively worded items

Across three arms, the positively worded items from the fourth block were presented differently (Appendix Table 2). In Arm 1, the block was preceded by a text warning: *“Please note that the next three questions look different and ‘most or all of the time’ is now a good thing.”* The wording and formatting (color and font size) were replicated from an earlier study [29]. In Arm 2, the response level order was reversed, with ‘most or all of the time’ at the top and ‘none of the time’ at the bottom. In Arm 3, no modifications were applied.

Statistical analysis

Respondent characteristics

Respondent characteristics were summarized using descriptive statistics. Differences across study arms were assessed using chi-square tests for categorical variables and one-way analysis of variance (ANOVA) for continuous variables, with Tukey’s honestly significant difference (HSD) tests for post hoc pairwise comparisons. Randomization was considered successful if no statistically significant differences were observed across arms for key matching variables, including age group, sex, education, region, residence type, and general health status.

Distributional properties

Response distributions, including floor and ceiling, were examined for all EQ-HWB items. Differences in distributions across study arms and between item variants were assessed using the chi-square test. Within-respondent differences for items with alternative formulations, conceptually matched groups (e.g., ‘hopelessness’ versus ‘nothing to look forward to’, ‘feeling excluded’ versus ‘feeling left out’ versus ‘feeling isolated’), or positive-negative item pairs were assessed using the Wilcoxon signed-rank test. Direct comparisons were conducted for:

- **Positively worded items** (‘feeling accepted’, ‘feeling good about oneself’, ‘doing things one wanted to’) **and their negatively worded counterparts** (‘feeling excluded’, ‘negative feelings about oneself’, ‘difficulties enjoyable activities’). Responses were visualized and cross-tabulated. Comparisons were conducted across study arms and between arms (Arm 1 versus 2, Arm 1 versus 3, and Arm 2 versus 3). For matched positive-negative pairs, positively worded items were reverse-coded for analysis to align score direction with their negatively worded counterparts.
- **Vision items with and without reference to aids.** Responses were visualized and examined according to use of vision aids (yes/no), presence of chronic eye conditions, frequency of aid use, and type of aid (e.g., reading or distance glasses). Agreement between items with and without reference to aids was assessed by calculating the proportion of identical and discordant responses. Among discordant responses, the direction of difference – whether greater difficulty with seeing was reported with or without reference to aids – was examined.

Response errors due to positively worded items

To identify potential response errors due to switching from negative to positive wording, responses to positively worded items were benchmarked against external measures, i.e., the happiness numeric rating scale, general well-being, and overall quality of life. Respondents were flagged as potentially inconsistent if they reported the least favorable response to positively worded items (e.g., ‘none of the time’ for ‘feeling good about oneself’) while simultaneously reporting high levels of happiness or excellent general well-being. Inconsistency was defined

in one direction, given the distribution of positively worded items and the expected impact of framing-related response error. Overlap across external measures was assessed, and the proportion of respondents flagged by at least one measure was estimated. To evaluate the impact of these potential inconsistencies, sensitivity analyses were conducted by excluding flagged respondents and examining whether expected monotonic patterns were restored (e.g., increasing mean happiness scores across response levels of positively worded items).

Convergent validity

Spearman's rank correlations were calculated for corresponding items to assess the extent to which they capture similar underlying constructs. This included comparisons between positively and negatively worded items, as well as corresponding item variants, grouped into conceptual domains: relationship-stigma ('feeling accepted', 'feeling excluded', 'feeling left out', and 'feeling isolated'), enjoyable activities ('doing things one wanted to do', 'difficulty with enjoyable activities', 'difficulty with doing rewarding things', 'difficulty with doing things one wanted to'), and self-worth ('feeling good', 'negative feelings', 'feeling bad about oneself', 'feeling like a failure') [2]. Additional corresponding pairs were also assessed (e.g., 'hopelessness' compared with 'nothing to look forward to'). Spearman's rho coefficients were interpreted as absolute values: negligible (< 0.10), weak (0.10 – 0.29), moderate (0.30 – 0.49), or strong (≥ 0.50) correlations [30].

Known-group validity

Known-group validity was assessed to evaluate the discriminatory power of alternative item formulations at both the item and instrument levels. At the item level, comparisons were conducted for positively and negatively worded items, and for 'hopelessness' and 'nothing to look forward to' items. Comparisons were examined across groups defined by general well-being, overall quality of life, happiness, self-esteem, chronic anxiety, and chronic depression. Additionally, known-group validity for vision items (with versus without aid reference) was assessed by comparing respondents with chronic eye disease and their healthy counterparts. At the instrument level, known-group validity was assessed using EQ-HWB LSS. This involved comparing the instrument's ability to distinguish between groups under alternative instrument versions: (1) replacing negatively worded items with their positively worded counterparts, and (2) replacing 'hopelessness' with 'nothing to look forward to' in the LSS. The resulting LSS were then transformed to a 0–100 scale for ease of interpretation, with higher scores reflecting better health and well-being, and compared across groups. The same grouping variables were used as in the item-level analyses, with additional groups defined by sociodemographic, care-related, and health characteristics, including education, caregiving status, perceived financial satisfaction, general health, chronic condition status, and EQ VAS. Group differences were assessed using independent samples t-tests, with Mann-Whitney U tests conducted as non-parametric sensitivity analyses. Cohen's d effect sizes, adjusted for sex and age, were calculated to quantify between-group differences and compare discriminative performance across items and instrument versions. Effect sizes were reported as absolute values, and interpreted as trivial (0.00 – 0.19), small (0.20 – 0.49), medium (0.50 – 0.79) or large (≥ 0.80) [30].

Analytical considerations

Analyses were exploratory in nature, and no a priori hypotheses were specified. All analyses were conducted using Stata/MP 19 (StataCorp LLC, 2025), with statistical significance set at $p < 0.05$.

Results

Respondent characteristics

Of 3,472 respondents recruited, 1,472 were excluded based on quality-control criteria or quota limits, primarily due to attention-check failures (67.9%) and exceeding quotas (23.9%). Other reasons included lack of consent (6.3%), selection of fictitious conditions (0.8%), failure of the year-of-birth consistency check (0.7%), and being under 18 years of age (0.3%). The final sample included 1,200 respondents (from $n=2,000$) across three arms with equal distribution, and randomization across arms was successful based on the matching criteria variables (all $p \geq 0.05$). The sample had a mean age of 49.3 ± 16.5 years, and 54.0% were female. The majority were married or partnered (61.3%), completed primary education only (43.6%), were employed (53.5%), and had lower income. Most (68.9%) had chronic conditions, and reported at least good levels of general health (61.5%), general well-being (60.7%), and general quality of life (61.7%). Overall, 763 (63.6%) respondents used vision aids, of whom 43.2% used reading and 30.4% used distance glasses (not mutually exclusive). Most users (78%) used them 6–7 days in the past week. Complete characteristics are reported in Appendix Table 3.

EQ-HWB item distribution

Except for the positively worded items, distributions of all EQ-HWB items and their variants (Blocks 1–3) are reported in Appendix Table 4 and visualized in Fig. 1. Items with the highest ceilings included ‘hearing’ (81.3%), ‘feeling unsafe’ (73.0%), ‘feeling like a failure’ (70.3%), and ‘getting around inside or outside’ (69.8%). Items with the lowest ceilings were ‘exhaustion’ (25.5%), ‘pain’ severity (34.1%), and ‘pain’ frequency (34.4%); these indicate the greatest reported difficulties among respondents. Distributions were largely similar across the three arms, with no statistically significant differences observed for most items. Exceptions were observed for ‘coping’ ($p=0.014$) and ‘difficulty with doing things one wanted to do’ ($p=0.033$), which showed differences across arms.

Distributions for positively vs. negatively worded items by arm

Figure 2 shows the distributions of the three positively worded items – ‘feeling accepted’, ‘feeling good about oneself’, and ‘doing things one wanted to’ – by study arm, alongside their corresponding negatively worded counterparts: ‘feeling excluded’, ‘negative feelings about oneself’, and ‘difficulty with enjoyable activities’. Negatively worded items showed higher ceilings (ranging from 50 to 52% for ‘negative feelings about oneself’ to 68–71% for ‘feeling excluded’), whereas positively worded items showed more dispersed distributions with lower ceilings (ranging from 17 to 26% for ‘feeling good about oneself’ to 25–39% for ‘feeling accepted’). The unmodified presentation (Arm 3) consistently showed lower ceilings compared with the modified arms for positively worded items. Within each arm, positively and negatively worded item pairs after reverse-coding showed significantly different response distributions ($p < 0.001$). Across arms, response distributions differed for ‘feeling good about oneself’ ($p=0.002$) and ‘doing things one wanted to’ ($p=0.001$), but not for ‘feeling accepted’ ($p=0.160$). Cross-tabulations are provided in Appendix Table 5.

Response errors on positively worded items

A proportion of respondents reported the least favorable responses to positively worded items (‘none of the time’) despite indicating high levels of happiness, general well-being, or quality of life, suggesting potential response errors. This pattern was more pronounced in the unmodified presentation arm (Arm 3), which consistently showed higher proportions of such responses compared with the modified arms (Arm 1: text warning prior to items; Arm 2: reversed response scale). The proportion of respondents flagged due to inconsistency with an external measure ranged from 3.5 to 4.5% in Arm 1, 3.3–4.0% in Arm 2, and 5.8–6.8% in Arm 3. As a robustness check, excluding

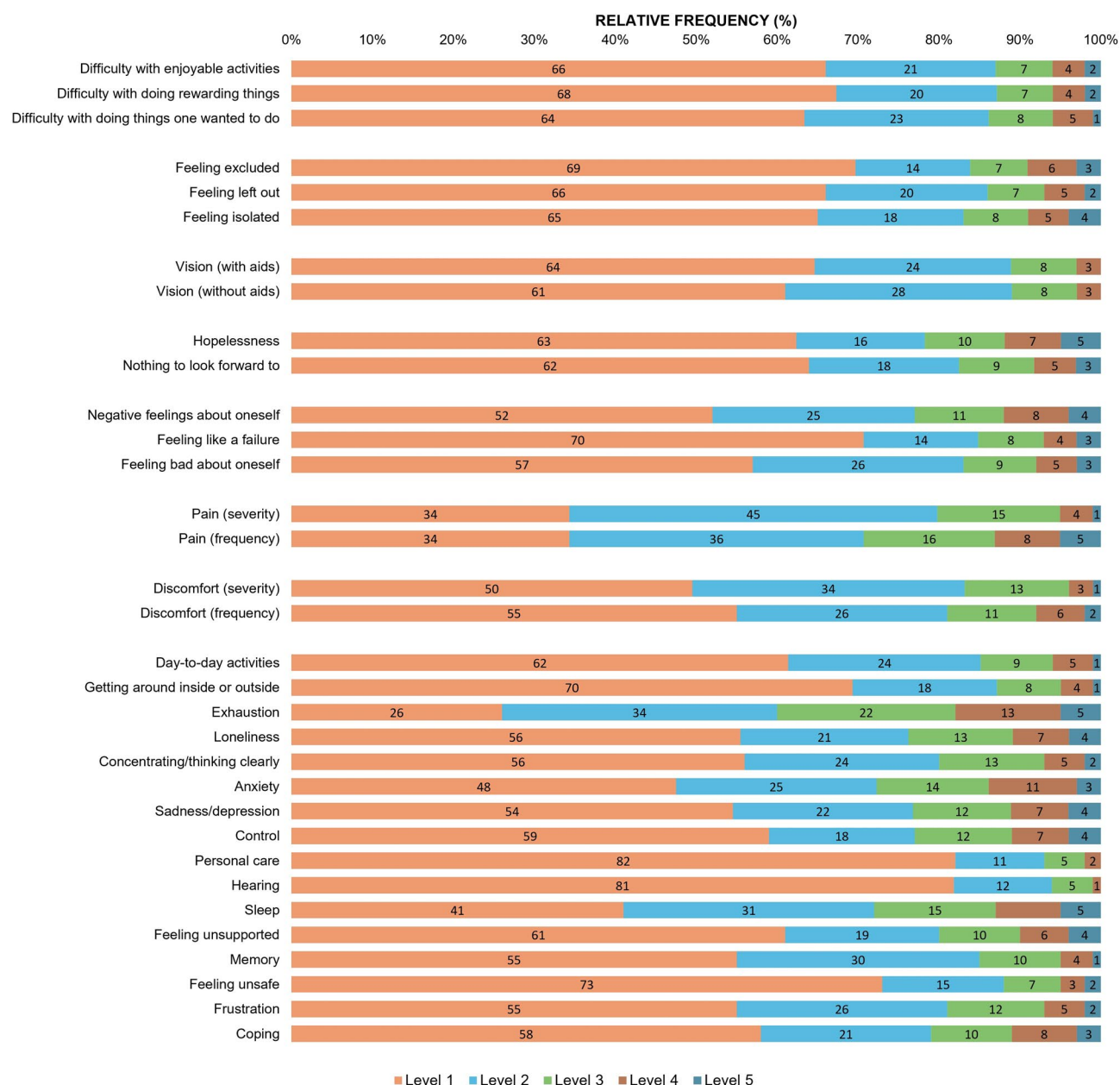


Fig. 1 Response distributions for EQ-HWB items and variants ($n=1,200$). Notes: (1) The figure excludes the positively worded items. (2) Selected items are arranged in pre-specified groups of two to three closely related items. (3) Percentages may not sum to 100% due to rounding. (4) Response categories with values less than 0.5% were omitted to improve readability. (5) Level 1 represents the most favorable response option for each item (e.g., ‘no difficulty’, ‘no pain’). Labels for response categories (levels 1–5) are provided in Appendix Table 4. (6) Wilcoxon signed-rank tests showed statistically significant within-respondent distributional differences for ‘difficulty with enjoyable activities’ vs. ‘difficulty with doing rewarding things’ ($p=0.005$), ‘difficulty with doing rewarding things’ vs. ‘difficulty with doing things one wanted to’ ($p<0.001$), ‘feeling left out’ vs. ‘feeling isolated’ ($p=0.004$), ‘hopelessness’ vs. ‘nothing to look forward to’ ($p<0.001$), ‘negative feelings about oneself’ vs. ‘feeling like a failure’ ($p<0.001$), ‘negative feelings about oneself’ vs. ‘feeling bad about oneself’ ($p<0.001$), ‘feeling like a failure’ vs. ‘feeling bad about oneself’ ($p<0.001$), and ‘pain’ severity vs. ‘pain’ frequency ($p<0.001$). No statistically significant differences were observed for ‘difficulty with enjoyable things’ vs. ‘difficulty with doing things one wanted to’ ($p=0.794$), ‘feeling excluded’ vs. ‘feeling left out’ ($p=0.947$), ‘feeling excluded’ vs. ‘feeling isolated’ ($p=0.083$), ‘vision’ with vs. without reference to aids ($p=0.053$), and ‘discomfort’ frequency vs. ‘discomfort’ severity ($p=0.130$).

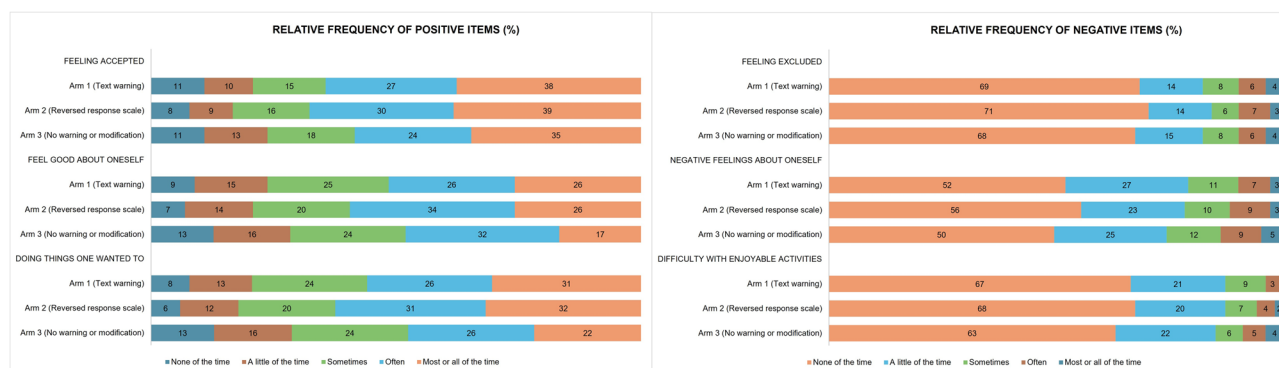


Fig. 2 Comparative response distributions for positively worded items and negatively worded counterparts. Notes: (1) Percentages may not sum to 100% due to rounding. (2) Positively worded items were presented differently across arms; arm 1: preceded by a text warning, arm 2: response level order reversed, arm 3: no text warning or any other modification. (3) Across the three presentation formats, response distributions for positively worded items were compared using chi-square tests. Statistically significant differences were observed for ‘feeling good about oneself’ ($p=0.002$) and ‘doing things one wanted to’ ($p=0.001$), but not for ‘feeling accepted’ ($p=0.160$). (4) Pairwise chi-square tests showed significant differences between arm 1 and arm 2 and between arm 2 and arm 3 for all three positively worded items (all $p<0.001$). Comparisons between arm 1 and arm 3 were not significant for ‘feeling accepted’ ($p=0.371$), but were significant for ‘feeling good about oneself’ ($p=0.009$) and ‘doing things one wanted to’ ($p=0.022$). (5) Matched comparisons between positively and negatively worded items within each arm were conducted using the Wilcoxon signed-rank test after aligning item direction (‘feeling accepted’ vs ‘feeling excluded’; ‘feeling good about oneself’ vs ‘negative feelings about oneself’; ‘doing things one wanted to’ vs ‘difficulty with enjoyable activities’); all differences were significant (all $p<0.001$)

flagged respondents restored the expected monotonic relationships between response levels and mean happiness scores for all items, with the lowest scores observed for ‘none of the time’ and progressively higher scores up to ‘most or all of the time’. Full results are presented in Appendix Table 6.

Convergence between positively and negatively worded items

Correlations between positively and negatively worded items were generally moderate across domains, with a few pairs reaching borderline strong correlations ($r=0.34$ – 0.52). Correlations tended to be slightly higher in Arms 1 and 2, where the positive wording was accompanied by a text warning or reversed response order, compared with Arm 3, where the items were presented without modification. Meanwhile, correlations among negatively worded items were strong ($r=0.63$ – 0.80). Results are presented in Appendix Table 7.

Known-group validity of positively and negatively worded items

Across all known groups, positively worded items presented under modified formats showed greater effect sizes than those in the unmodified format, with the largest discrimination observed for groups defined by happiness and self-esteem (Table 1; see Appendix Table 8 for full details). Except for the self-esteem subgroup, corresponding negatively worded items showed greater effect sizes than their positively worded counterparts. At the instrument level, replacing negatively worded items with positively worded items slightly improved discriminatory power across all known groups and study arms (Appendix Table 9).

Hopelessness versus nothing to look forward to items

‘Hopelessness’ and ‘nothing to look forward to’ showed similar distributions, with ceilings of 62.5% and 65.3%, respectively (Fig. 1), and were strongly correlated ($r=0.69$, Appendix Table 7). In known-group validity analyses, hopelessness consistently demonstrated larger effect sizes than nothing to look forward to (Table 2). However,

Table 1 Known-group validity of positively worded items and negatively worded counterparts (Cohen's d effect sizes; full results in Appendix Table 7)

Known groups	Arm 1			Arm 2			Arm 3			Positively worded items			Feeling good about oneself			Doing things one wanted to			Negatively worded items			n	Difficulty with Enjoyable activities
	Arm 1			Arm 2			Arm 3			Feeling accepted			Feeling excluded			Feeling good about oneself			Feeling excluded				
	n	n	n	n	n	n	n	n	n	Arm 1	Arm 2	Arm 3	Arm 1	Arm 2	Arm 3	Arm 1	Arm 2	Arm 3	Arm 1	Arm 2	Arm 3		
General well-being																							
Fair/poor	147	154	171	0.605	0.597	0.427	0.607	0.628	0.512	0.642	0.664	0.538	0.693	0.685	0.659	(M)	(M)	(M)	472	0.659	(M)	0.659	(M)
Excellent/(very) good	253	246	229	(M)	(M)	(S)	(M)	(M)	(M)	(M)	(M)	(M)	(M)	(M)	(M)	(M)	(M)	(M)	728	(M)	(M)	728	(M)
General quality of life																							
Fair/poor	155	151	154	0.603	0.521	0.382	0.637	0.564	0.557	0.679	0.465	0.541	0.723	0.660	0.630	(M)	(M)	(M)	460	0.630	(M)	0.630	(M)
Excellent/(very) good	245	249	246	(M)	(M)	(S)	(M)	(M)	(M)	(M)	(S)	(M)	(M)	(M)	(M)	(M)	(M)	(M)	740	(M)	(S)	740	(M)
Happiness 0–10 NRS																							
Not happy (0–5)	96	97	109	1.082	1.099	0.686	1.122	1.273	0.938	1.035	1.105	0.820	1.460	1.203	0.990	(L)	(L)	(L)	302	0.990	(L)	0.990	(L)
Happy (6–10)	304	303	291	(L)	(L)	(M)	(L)	(L)	(L)	(L)	(L)	(L)	(L)	(L)	(L)	(L)	(L)	(L)	898	(L)	(L)	898	(L)
Self-esteem^b																							
Below adequate (0–15)	141	117	136	0.947	1.368	1.012	1.076	1.445	1.210	0.937	1.166	0.960	1.202	1.324	0.955	(L)	(L)	(L)	394	0.955	(L)	0.955	(L)
Adequate (> 15)	259	283	264	(L)	(L)	(L)	(L)	(L)	(L)	(L)	(L)	(L)	(L)	(L)	(L)	(L)	(L)	(L)	806	(L)	(L)	806	(L)
Chronic anxiety																							
Yes	41	42	46	0.837	0.720	0.394	0.904	0.930	0.708	0.836	0.283	0.797	1.685	1.633	1.462	(L)	(L)	(L)	129	1.462	(L)	1.462	(L)
Healthy	105	93	106	(L)	(M)	(S)	(L)	(L)	(M)	(L)	(S)	(M)	(L)	(S)	(M)	(L)	(S)	(M)	304	(L)	(S)	304	(L)
Chronic depression																							
Yes	31	31	35	0.927	0.723	0.621	0.848	1.125	0.833	0.864	0.490	0.908	1.969	1.903	1.823	(L)	(L)	(L)	97	1.823	(L)	1.823	(L)
Healthy	105	93	106	(L)	(L)	(M)	(L)	(L)	(L)	(L)	(S)	(L)	(L)	(S)	(L)	(L)	(S)	(L)	304	(L)	(S)	304	(L)

Abbrev. L: large effect size, M: medium effect size, NRS: numeric rating scale

General notes: Cohen's d values are reported in absolute terms and interpreted as: 0.00–0.19 trivial, 0.20–0.49 small, 0.50–0.79 medium, and ≥0.80 large. All comparisons were adjusted for age and sex

Specific notes. (a) Presented differently across arms; arm 1: preceded by a text warning, arm 2: response level order reversed, arm 3: no text warning or any other modification. (b) Measured using the Rosenberg Self-Esteem Scale

Table 2 Known-group validity of ‘hopelessness’ and ‘nothing to look forward to’ items

Known group	n	%	Hopelessness		Nothing to look forward to	
			Mean (SD)	Effect size	Mean (SD)	Effect size
General well-being						
Fair/poor	472	39%	2.29 (1.36)	0.932 (L)	2.04 (1.25)	0.778 (M)
Excellent/(very) good	728	61%	1.39 (0.83)		1.34 (0.72)	
General quality of life						
Fair/poor	460	38%	2.27 (1.36)	0.866 (L)	2.05 (1.23)	0.764 (M)
Excellent/(very) good	740	62%	1.42 (0.86)		1.35 (0.75)	
Happiness 0–10 NRS						
Not happy (0–5)	302	25%	2.90 (1.35)	1.641 (L)	2.51 (1.30)	1.360 (L)
Happy (6–10)	898	75%	1.36 (0.76)		1.32 (0.68)	
Self-esteem^a						
Below adequate (score 0–15)	394	33%	2.63 (1.34)	1.330 (L)	2.32 (1.28)	1.192 (L)
Adequate (score > 15)	806	67%	1.31 (0.74)		1.27 (0.62)	
Chronic anxiety						
Yes	129	30%	3.09 (1.41)	1.768 (L)	2.67 (1.40)	1.629 (L)
Healthy	304	70%	1.35 (0.79)		1.22 (0.59)	
Chronic depression						
Yes	97	24%	3.34 (1.32)	2.147 (L)	2.84 (1.37)	1.943 (L)
Healthy	304	76%	1.35 (0.79)		1.22 (0.59)	

Abbrev. L: large effect size, M: medium effect size, NRS: numeric rating scale

General notes. (1) Known-group validity was assessed using t-tests; results were robust to non-parametric sensitivity analyses using the Mann-Whitney U test. All *p*-values were < 0.001. (2) Item scores (1–5) were aligned such that higher values indicate greater difficulty. (3) Effect size: Cohen's *d* values are reported in absolute terms and interpreted as: 0.00–0.19 trivial, 0.20–0.49 small, 0.50–0.79 medium, and ≥ 0.80 large. All known-group comparisons were adjusted for age and sex

Specific notes. (a) Measured using the Rosenberg Self-Esteem Scale

at the instrument level, substituting one for the other resulted in similar effect sizes for the EQ-HWB across all known groups (Appendix Table 10).

Vision items with versus without reference to aids

Figure 3 shows the distribution of ‘vision’ items. Among vision aid users (*n* = 763), distributions differed slightly, with a higher ceiling observed for the item with reference to aids (52.3%) compared with without reference (47.6%). A similar pattern was observed across subgroups (by chronic eye disease, type, and frequency of aid use), with consistently higher ceilings for the item with reference to aids than without, and comparable differences in percentage points. The lowest ceilings were observed among respondents with chronic eye disease (*n* = 91), at 20.9% with aid reference and 24.2% without. Significant differences between item versions were observed among vision aid users (*p* = 0.014), reading glasses users (*p* = 0.003), and those using aids ≤ 5 days per week (*p* = 0.034), but not among non-users, respondents with chronic eye disease, distance glasses users, or those using aids 6–7 days per week.

Overall, 85% of respondents gave identical responses to both items. Among discordant responses, more respondents reported greater seeing difficulty without reference to aids (8.6%) than with reference (6.4%) (Appendix Table 11). Among vision aid users, 81.5% responded identically, with 11.1% reporting greater difficulty without reference compared with 7.3% with reference. Both item versions demonstrated similar discriminatory performance in distinguishing respondents with chronic eye disease from healthy counterparts, with large effect sizes (Cohen's *d* = 1.79 with aid reference vs. 1.76 without).

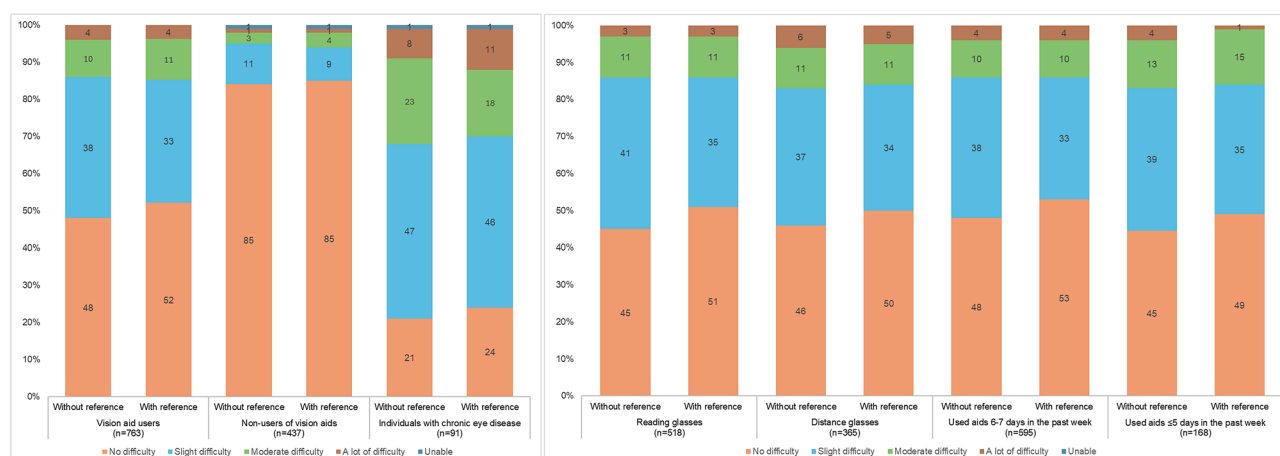


Fig. 3 Vision item responses with and without reference to aids. Notes: (1) Percentages may not sum to 100% due to rounding; (2) Response categories with values less than 0.5% were omitted to improve readability. (3) Paired comparison of item reference (with vs. without reference to aids): • Vision aid use: users of vision aids ($n=763$), $p=0.014$; Non-users of vision aids ($n=437$), $p=0.619$; • Individuals with chronic eye disease ($n=91$), $p=0.654$; • Reading glasses users ($n=518$), $p=0.003$; • Distance glasses users ($n=365$), $p=0.073$; • Frequency of use: 6–7 days ($n=595$), $p=0.198$; 5 days or less ($n=168$), $p=0.034$

Discussion

This study provides one of the most extensive quantitative evaluations of EQ-HWB long form modifications to date. Findings show substantial differences in distributions between positively and negatively worded items. While these items may appear antonymous, they are not strictly equivalent. For example, not feeling accepted does not necessarily imply feeling rejected [22]; therefore, responses are not expected to mirror each other. Furthermore, the shift from negatively to positively worded items appeared to confuse some respondents, even when placed last in the questionnaire, as it requires a change in response orientation. The mitigation strategies reduced, but did not eliminate, response errors. These approaches show promise; however, the presentation formats for the positively phrased items evaluated in this study were developed *de novo* - qualitative research would be valuable to inform further refinement. Combining approaches (e.g., a text warning with reversed response options) may offer additional improvements. Another approach could involve real-time confirmation prompts following extreme responses (e.g., “you selected X, which means Y – could you please confirm?”). In paper-based administration, alternative approaches would be required, as interactive checks are not feasible. While positively worded items showed potential value in improving instrument validity, this benefit may be offset by the introduction of response errors.

Regarding alternative item performance, a slightly higher proportion of respondents reported greater difficulty on the ‘vision’ item without reference to aids compared with the version including such reference. Differences were observed among vision aid users, but not among those with chronic eye conditions, suggesting that respondents accounted for their use of aids when answering. These findings may indicate that explicitly referencing the use of aids has limited added value in a general population sample. However, this should not be interpreted as evidence that this reference is unnecessary. The EQ-HWB was informed by the Wilson and Cleary model [1], which recognizes the role of individuals’ perceptions and experiences in self-reported functioning. For many individuals, aids such as glasses form part of usual functioning and should therefore be considered when assessing vision-related functioning. While previous qualitative evidence has suggested possible confusion around the wording of the use-of-aids reference, this appears to relate to phrasing rather than to the relevance of the reference itself, indicating the need for further refinement [7, 22]. The Hungarian version also showed similar issues in the cognitive debriefing following the initial translation, which were addressed through minor rephrasing (e.g.,

replacing ‘using for example glasses or contact lenses if you normally use them’ with ‘if you used, for example, glasses or contact lenses, please consider their use when answering’); this revised wording was used in the present study. Further testing in populations with visual impairment is warranted to better understand the impact of including a reference to aids in the vision item. For the hopelessness wording comparison, the ‘hopelessness’ item demonstrated stronger known-group validity than ‘nothing to look forward to’, while showing similar distributions and comparable effects at the instrument level, suggesting that this substitution performs well.

This exploratory study has several limitations. The study was conducted in Hungarian, using an online general-population panel in Hungary; as the EQ-HWB is intended for international use, framing and response-order effects may differ across languages and cultural contexts. The EQ-HWB items and their variants were administered in a fixed order, which may have introduced order effects such as anchoring or priming. For example, respondents were consistently presented with the vision item including reference to aids before the version without such reference, potentially influencing subsequent responses. Similarly, the sequencing of negatively worded items before positively worded ones may have affected respondents’ mindset and contributed to framing-related response patterns. Items were also administered in blocks and in an order that differed from the full instrument. In addition, item order differed across arms for some items, which may have influenced responses. The inclusion of multiple item variants, additional instruments, and supplementary questions may also have increased respondent burden, potentially affecting response quality despite the implemented data quality checks. Some included instruments, such as the Rosenberg Self-Esteem Scale, also used mixed item framing (positive and negative), making it unclear whether the observed confusion was specific to the EQ-HWB item framing switch. Furthermore, because our definition of potential response errors focused on least-favorable responses to positively worded items despite favorable external indicators, other possible patterns of response inconsistency may not have been captured. Despite these limitations, the findings inform key design decisions for the EQ-HWB long form, particularly in balancing gains in psychometric performance with the risk of response errors. Further validation in other administration modes would help assess the robustness of these findings, and qualitative studies would be valuable to inform refinement.

Conclusions

Positively worded items in the EQ-HWB improved item distributional properties and instrument-level known-group validity but increased response errors from negative-to-positive wording switches. Alternative presentation formats reduced but did not eliminate these errors. The ‘hopelessness’ item outperformed ‘nothing to look forward to.’ Including a reference to aids in the vision item had only a marginal effect in a general population sample. These findings contribute to evidence informing decisions on the finalization of the EQ-HWB long form.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1186/s12955-026-02621-w>.

Acknowledgements The authors thank all study participants for their involvement in this study.

Author contributions All authors (S.P., C.B., C.M., T.P., A.B., E.S., and F.R.) contributed to the conception and design of the study. S.P. performed the data analysis and drafted the original manuscript. All authors contributed to the interpretation of the findings, reviewed and edited the manuscript, and approved the final version for submission.

Funding This study was supported by the EuroQol Research Foundation (2323-RA).

Data availability The data supporting the findings of this study are available from F.R. upon reasonable request.

Declarations

Ethics approval and consent to participate Ethics approval was granted by the Research Ethics Committee of Corvinus University of Budapest (KRH/284/2024). All respondents provided informed consent prior to participation. All procedures involving human participants were conducted in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Declaration of Helsinki and its later amendments, or comparable ethical standards.

Role of the funder 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; and decision to submit the manuscript for publication.

Authorship confirmation All authors certify that they meet the ICMJE criteria for authorship.

Competing interests Stevanus Pangestu and Fanni Rencz are employed by the EuroQol Research Foundation. Cate Bailey, Clara Mukuria, Tessa Peasgood, and Fanni Rencz are members of the EuroQol Group. All authors have received grants from EuroQol for work outside of the scope of this study. Views expressed in the article are those of the authors and are not necessarily those of the EuroQol Research Foundation.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

1. Brazier J, et al. The EQ-HWB: Overview of the Development of a Measure of Health and Wellbeing and Key Results. *Value Health*. 2022;25(4):482–91.
2. Peasgood T, et al. Developing a New Generic Health and Wellbeing Measure: Psychometric Survey Results for the EQ-HWB. *Value Health*. 2022;25(4):525–33.
3. Bailey C, et al. The Psychometric Performance of the EQ Health and Wellbeing in Caregivers of Persons Living With Dementia. *Value Health*. 2025;28(8):1221–30.
4. Bailey C, et al. Validity and Reliability of the EQ-HWB in Parents of Children With Health Conditions. *Value Health*. 2025;28(9):1370–9.
5. Bailey C, et al. The performance of the EQ-HWB-S as a measure of quality-of-life of caregivers in families that have experienced adverse events. *Eur J Health Econ*. 2025;26(1):7–21.
6. Bailey C, et al. The psychometric performance of the EQ-HWB-9 for measuring health and wellbeing in a general population sample from Australia and New Zealand. *Qual Life Res*. 2025;34(12):3707–19.
7. Engel L, et al. Appropriateness of the EQ-HWB for Use in Residential Aged Care: A Proxy Perspective. *Patient*. 2024;17(6):673–83.
8. Kinchin I, Engel L, Rencz F. A Comparative Study of Health and Well-Being Measures in Ireland Using EQ Health and Wellbeing (EQ-HWB) and its Short Version, EQ-5D-5L, and ICEpop Capability Measure for Adults (ICECAP-A). *Value Health*. 2025;28(8):1268–79.
9. Kuharic M, et al. Comparison of the EQ-HWB and EQ-HWB-S With Other Preference-Based Measures Among United States Informal Caregivers. *Value Health*. 2024;27(7):967–77.
10. Kuharić M, et al. The Measurement Properties of the EQ-HWB and the EQ-HWB-S in Italian Population: A Comparative Study With EQ-5D-5L. *Value Health*. 2024;27(7):955–66.
11. Keetharuth A, et al. EQ Health and Wellbeing (EQ-HWB): a psychometric assessment across 6 conditions and the general population in the United Kingdom. *Value Health*. 2025.
12. Lee P, et al. Exploring the Comparability Between EQ-5D and the EQ Health and Wellbeing in the General Australian Population. *Value Health*. 2024;27(4):508–17.
13. Long C, Mao Z, Yang Z. A Head-to-Head Comparison of EQ-HWB and EQ-5D-5L in Patients, Carers, and General Public in China. *Value Health*. 2024;27(7):848–56.

14. McDool E, Mukuria C, Peasgood T. Psychometric Performance of the EQ Health and Wellbeing Short in a United Kingdom Population Sample. *Value Health*. 2024;27(9):1215–24.
15. McDool E, et al. Evaluating the validity of the EQ-HWB-9 in a large UK general population sample. *Value Health*. 2026.
16. Nasser SC, et al. Frequency and severity response scales for pain and discomfort: psychometric insights from EQ-HWB. *Qual Life Res*. 2025;34(8):2259–68.
17. Pangestu S, et al. The Psychometric Properties of the EQ-HWB and EQ-HWB-S in Patients With Breast Cancer: A Comparative Analysis With EQ-5D-5L, FACT-8D, and SWEMWBS. *Value Health*. 2025;28(3):449–59.
18. Xu RH, Yang C, Rencz F. Exploring recall periods for EQ-5D-5L and EQ-HWB-9: a Hong Kong population study. *Value Health*. 2026.
19. Liao M, et al. Validity and responsiveness of the EQ-5D-5L, EQ-HWB and EQ-HWB-9 to measure health and wellbeing impact of heatwaves among older adults. *Health Qual Life Outcomes*. 2026.
20. Xu RH, Rencz F. Psychometric Properties of the Modified EQ Health and Wellbeing and its Shorter Version in a General Chinese Population: A Comparative Analysis With EQ-5D-5L, SF-6Dv2, ICEpop CAPability measure for Adults, Short Version of Warwick-Edinburgh Mental Wellbeing Scale, and World Health Organization Well-Being Index 5-Item. *Value Health*. 2025;28(8):1241–52.
21. Akanksha A, et al. *Testing possible modifications to the EQ-HWB in a sample of people with chronic conditions*, in *5th EuroQol Early Career Researcher Meeting 2025*. Spain: Barcelona; 2025.
22. Bailey C, Peasgood T. Content validity of the EQ-HWB in caregivers of children with health conditions. *Patient*. 2025.
23. Metyovinyi Z, et al. Qualitative testing of the EQ-HWB-9 with a ‘blank’ EQ-HWB VAS and EQ-5D-5L with skin bolts in a rare genetic skin disease population. *Qual Life Res*. 2026;35(3):69.
24. Bató A, et al. Reference group use and the usefulness of age- and self-comparative EQ-5D-5L modifications: a mixed-methods study. 42nd EuroQol Plenary 2025. Austria: Vienna; 2025.
25. Herdman M, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011;20(10):1727–36.
26. Rosenberg M. Rosenberg self-esteem scale. *J Relig Health*. 1965.
27. Prawitz A, et al. InCharge financial distress/financial well-being scale: Development, administration, and score interpretation. *J Financial Couns Plann*. 2006;17(1).
28. Carlton J, et al. Generation, Selection, and Face Validation of Items for a New Generic Measure of Quality of Life: The EQ-HWB. *Value Health*. 2022;25(4):512–24.
29. Thai T, et al. Comparing the original and modified versions of the EQ-HWB-9: a psychometric evaluation in an Australian general population sample. 42nd EuroQol Plenary 2025. Austria: Vienna; 2025.
30. Cohen J. A power primer. *Psychol Bull*. 1992;112(1):155–9.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Stevanus Pangestu^{1,2}  · Cate Bailey³ · Clara Mukuria⁴ · Tessa Peasgood⁵ · Alex Bató⁶ · Eszter Szlávicz^{7,8} · Fanni Rencz^{1,6}

✉ Stevanus Pangestu
pangestu@euroqol.org

¹ EuroQol Research Foundation, Marten Meesweg 107, Rotterdam 3068 AV, Netherlands

² Department of Public Health, Erasmus University Medical Center, Rotterdam, Netherlands

³ School of Population and Global Health, The University of Melbourne, Carlton, Australia

⁴ CM HE Consulting, Sheffield, UK

⁵ School of Medicine and Population Health, University of Sheffield, Sheffield, UK

⁶ Department of Health Policy, Corvinus University of Budapest, Budapest, Hungary

⁷ Department of Dermatology, Venereology and Oncodermatology, The University of Pécs Medical School, Pécs, Hungary

⁸ Mental Health and Psychology Centre, Archiepiscopal College of Veszprém, Veszprém, Hungary