Testing possible modifications to the EQ-HWB in a sample of people with Chronic Conditions

Akanksha Akanksha[1], Tessa Peasgood[2], Richard Norman[3], Deborah Street[1], Rosalie Viney[1], Brendan Mulhern[1]

[1] Centre for Health Economics Research and Evaluation, University of Technology Sydney, Sydney, NSW, Australia
[2] School of Health and Related Research, University of Sheffield, Sheffield, England, UK
[3] School of Population Health, Curtin University, Perth, WA, Australia

Abstract

Background: The EQ Health and Wellbeing (EQ-HWB) is a generic outcome measure developed by the EuroQol group to assess health and well-being across different sectors. The 25 (EQ-HWB) and 9-item (EQ-HWB-S) versions are currently designated as experimental instruments. Although preliminary evidence supports the EQ-HWB's measurement properties, item modifications could enhance its performance, clarity, and user acceptability. A notable concern involves the three positively framed items included in the long form version. These items might lead to response inconsistencies due to completion error and differing interpretations of response options when switching between positively and negatively framed items. This study aims to assess the performance of negatively framed versions of the three positively framed items. Additionally, we investigate whether 'flagging' the switch to the positively framed EQ-HWB items reduces the likelihood of response inconsistencies by the respondents.

Methods: This study utilized data from an online longitudinal survey conducted in Australia that recruited 812 adults (18+) at baseline and 495 at follow-up, 14 days apart, across four chronic condition groups (diabetes, respiratory conditions, heart conditions, and anxiety/depression). We tested nine modified items (three for each of the positively framed items). Measurement properties evaluated included item distribution, convergence, test-retest reliability, and known-group differences using Cohen's D, and the Wilcoxon Rank-Sum Test. Confirmatory Factor Analysis (CFA) was used to assess the domain structure of each construct measured by the positive items, and their negative counterparts. A chi-square test was conducted to assess whether the flagging of positively framed items in the questionnaire influenced response patterns.

Results: The distribution of the negatively framed test items was similar to that of their positive counterparts; however, four test items (three of which were the modifications for the item- 'doing things one wants to do') out of nine had a response distribution of <5% at the most severe level. Test items showed moderate correlations with their positive counterparts (0.40 to 0.55), while stronger correlations (>0.70) were observed for other EQHWB items, such as lonely, unsupported, sad/depressed, look forward and cope. Weighted Cohen's kappa values were slightly higher for the test items (0.54 to 0.64), than the original positive items (0.45–0.51). Known group differences revealed slightly better performance for the negatively framed test items, (ES range 0.47 to 0.73) compared to the positively framed items (ES 0.45 to 0.62). In the comorbidities group, the effect size for the negatively framed items ranged from 0.24 to 0.54, and the effect size for the positively framed items was 0.17 to 0.26. CFA demonstrated significant factor loadings, with a root mean square error of approximation range of 0.03–0.05, and Tucker-Lewis index and comparative fit index values above 0.99. Notifying respondents before the positively framed items impacted responses to the "*feeling accepted by others*" item (χ^2= 15.2; p=0.004).

Conclusion: The measurement properties of the negatively framed items were comparable, and in some cases better, than the original items. This indicates their potential to enhance measurement consistency in chronic condition populations. The evidence from this quantitative investigation can be used alongside other international mixed methods studies to inform any future changes to the item content of the EQ-HWB.

Funding: This study was funded by grants from the Australian Research Council and EuroQol Research Foundation. The views are those of the authors.

1. Background

Generic outcome measures play a crucial role in assessing quality of life (QoL), enabling comparisons across different conditions, populations, and sectors, essential for economic evaluations in healthcare. Recently, there has been growing interest in integrating a broader wellbeing framework into economic evaluations, alongside traditional measures of health-related quality of life (HRQoL). This interest arises from the recognition that for certain interventions—such as long-term care, social care, and end-of-life care—the goal extends beyond simply improving health or prolonging life; the focus is also on enhancing overall wellbeing (Brazier & Tsuchiya, 2015).

A notable recent addition is the EQ Health and Wellbeing (EQ-HWB), a newly developed generic instrument designed by the EuroQol group to assess the impact of health, social care, and informal caregiving on QoL (Brazier et., 2022; Peasgood et al., 2022). This measure represents a significant step forward in capturing both health and broader wellbeing constructs. The EQ-HWB has two versions: a 25-item profile measure and a 9-item preference weighted measure (PWM), the EQ-HWB Short (EQ-HWB-S) (Brazier et al., 2022). These measures were developed in an international study conducted across six countries: Argentina, Australia, China, Germany, the United Kingdom, and the United States (Brazier et., 2022; Peasgood et al., 2022). The EQ-HWB has the potential for widespread use across a range of settings, for example in clinical practice and population health surveys. As a PWM, the EQ-HWB-S will be used for economic evaluation across health, social care, and public health sectors (Kuharic et al., 2024).

As a newly developed instrument, both versions of the EQ-HWB are still considered "experimental" (in the EuroQol Group's IP status hierarchy). Recent evidence supports the face, content, convergent and known group validity of both the long and short versions (Masutti et al., 2024; Kuharic et al., 2024; McDool et al., 2024), but concerns have been raised about the interpretation of the three positively framed items included in the EQ-HWB appearing alongside negatively framed items. This is because the use of positively framed items alongside negatively framed items, while generally more amenable to respondents, may introduce response inconsistencies. Wording effects linked to the item direction can result in low correlation between positive and negative items (Ponce et al., 2022; Sonderen et al., 2013). It has also been shown that reverse-worded items often fail to load onto the

same factor as positively worded items, likely reflecting differences in cognitive processing, or response errors based on the ordering of positive and negative items, rather than true variance in the underlying trait (Salazar, 2015). These inconsistencies can compromise the reliability and validity of scales, particularly in large-scale surveys or interventions (Swain et al., 2008). While positively framed items are often seen favourably by respondents, their exclusive use or careful testing of mixed-item formats is recommended to minimize measurement error (Ebesutani et al., 2012).

This issue reflects a broader challenge in survey design and has implications for valuation. For instance, the inclusion of a positively framed item in the SF-6Dv1 descriptive system influenced valuation, resulting in inconsistencies (Brazier et al., 2020). This challenge is exacerbated by the need for respondents to imagine health states where, as an example, for multiple domains, responses such as "most of the time" indicate a poor HRQoL, while for one domain, "most of the time" reflects a high HRQoL. Such contrasts can create cognitive dissonance or misinterpretation, complicating the valuation process. While combining positive and negative items is often intended to reduce acquiescence bias—where respondents tend to agree with statements regardless of their content—research suggests that this approach may not always be effective (Sauro and Lewis, 2011). Instead, it can sometimes introduce inconsistencies, particularly when respondents struggle to differentiate between negatively and positively worded items (Salazar, 2015).

The recognition of potential response inconsistencies arising from the use of positively framed items in the EQ-HWB has prompted the exploration of two potential solutions. One approach is to frame all items consistently in one direction, to reduce respondent confusion and improve comparability across items. The alternative approach focuses on minimizing the chance of errors in completion through careful survey and questionnaire design, such as flagging shifts in framing to alert respondents. However, it is important to note that while this second approach may mitigate errors during completion, it does not fully address the challenges these items pose in the context of valuation. Building on these considerations, this study aims are to:

1. Evaluate the psychometric characteristics of negatively framed versions of the three positively framed items included in the EQ-HWB.

2. Assess whether flagging the switch to the positively framed EQ-HWB items reduces the likelihood of respondents misinterpreting the item response scales.

2. Methods

2.1 Data description

This study utilized longitudinal data collected between May and June 2024 as part of an ongoing project funded by the Australian Research Council (ARC) (DP210102021). An online survey, administered through the Decipher platform, recruited adults aged 18 and older across four health condition groups: diabetes, chronic respiratory conditions (asthma and COPD), chronic heart conditions, and anxiety/depression. Recruitment was facilitated by the panel company Pureprofile, which accessed its database of consenting, registered panel members, categorized by health condition.

At baseline and follow-up (two-weeks apart), respondents completed a series of generic and condition-specific QoL instruments. The survey was tested in a pilot phase, and routine quality checks were implemented for both the initial and follow-up surveys. Screening questions and data quality assessments were embedded throughout, alongside additional checks from the panel company, such as repeated questions about health status and age. All potential respondents had access to detailed study information prior to participation, and informed consent was obtained prior to respondents completing the survey. Ethics approval was granted by the UTS Human Research Ethics Committee (HREC approval number ETH24-9036).

2.2 Additional items and flags developed and included

To assess Aim 1, 50% of respondents were presented with an additional set of nine negatively framed questions (three modified questions each for the three positively framed EQ-HWB items). The full set of nine negatively framed modified test questions, along with the corresponding positively framed questions are provided in Table 1. These were designed to mirror the original positively framed items, providing an alternative version to assess the impact of negatively framed questions on response consistency and bias. These additional questions were placed at the end of the original EQ-HWB questionnaire, ensuring they were presented after the participants had already completed the full set of 25 items.

To assess Aim 2, 50% of respondents were randomly selected to receive a notification before completing the three positively framed items in the EQ-HWB. The notification, which appeared prior to questions Q19–Q21(*"Did you feel accepted by others?* (*e.g., feeling like you could be yourself and belonged)*"; *"Did you feel good about yourself?"*; and *"Could you do the things you wanted to do?")*, was intended to highlight the distinction in wording, helping participants interpret the questions more clearly and evaluate if the notification improves response accuracy. The message read: "Please note that the next three questions look different, and 'most or all of the time' is now a good thing."

2.3 Data analysis – Aim 1

The following analyses were conducted to test the modified items. All analyses were carried out using Stata version 18 (StataCorp, 2023)

2.3.1 Item distribution and correlation

The distribution of responses across item categories was examined to identify any potential issues. Items were considered to be at risk of ceiling effects if more than 70% of respondents selected the top category. Additionally, a full correlation matrix was generated using Spearman's rank correlations to assess the relationships between items both within and across dimensions. Items with correlations <0.5 within their respective subdomains were flagged as they may indicate insufficient alignment with the construct being measured, suggesting that such items may not adequately capture the intended domain. Conversely, items with correlations >0.7 with items from other subdomains were flagged due to the potential for redundancy, as such patterns may indicate overlap with other constructs, thereby compromising the measure's discriminant validity.

2.3.2 Test-retest reliability

Respondents who reported no changes in health, social circumstances, or overall quality of life (measured by questions asking about any change in each since the last survey) during the 14-day recall period were included in the analysis. Test-retest reliability of the EQ-HWB was assessed using weighted Cohen's kappa statistics, which assess the level of agreement between responses (Cohen, 1968). A weighted kappa closer to 1 indicates strong agreement, while values closer to 0 suggest minimal agreement beyond chance. The level of agreement

is classified as <0.00 (poor); 0.00-0.20 (slight); 0.21-0.40 (fair); 0.41-0.60 (moderate); 0.61-0.80 (substantial); 0.81-1.00 (almost perfect) (Landis & Koch, 1977).

2.3.4 Known-groups differences

The ability of items to discriminate between groups with known differences was assessed using Cohen's D and Wilcoxon Rank-Sum test. Cohen's D was calculated as the mean difference between groups divided by the pooled standard deviation, with effect sizes classified as small (0.2 to 0.49), medium (0.5 to 0.79), and large (≥0.8) (Cohen, 1992). The groups were defined as follows: Group 1 (VAS score ≥80) and Group 2 (VAS score <80), and Group 1 (reported other health conditions) and Group 2 (no other health conditions reported). The grouping by other health conditions explored whether the presence of any co-occurring conditions—regardless of severity—contributes to differences in health outcomes.

Wilcoxon Rank-Sum test was conducted to account for the ordinal nature of the EQ-HWB 5-point scale. The Wilcoxon rank-sum test compares two independent groups to determine if they come from the same distribution (Harris & Hardin, 2013). This non-parametric test is suitable for ordinal data which is skewed and provides a more accurate assessment of differences between groups.

2.3.4 Confirmatory factor analysis

Confirmatory factor analysis (CFA) was conducted to evaluate domain structure. CFA was performed separately for each domain (i.e. three domains were tested corresponding to each set of positive and negatively framed twinned items) to assess how well the observed items aligned with the factor structure. This approach allowed for the testing of the instrument's construct validity by confirming whether the items within each domain were consistently measuring the intended latent constructs. CFA models were judged based on the following fit indices: root mean square error of approximation (RMSEA) <0.6 taken as good, comparative fit index (CFI) >0.95 taken as good, and Tucker Lewis index (TLI) >0.95 taken as good (Hu & Bentler, 1999).

2.4 Data analysis – Aim 2

To evaluate the impact of the notification about the framing change (positive vs. negative), the Chi-Square test was used to compare the response distributions for items Q19–21

between the group that received the notification and the group that did not. A significant result would be consistent with the notification having influenced how respondents interpreted and responded to these items.

3. Results

3.1 Descriptive Statistics

The study sample consisted of 812 respondents at baseline, with 495 (61%) completing the follow-up. Of the total sample, 425 (52%) were female and 386 (48%) were male. The majority, 674 (83%), were born in Australia. Respondents were evenly distributed across the four health condition groups. Socio-demographic characteristics of the sample are detailed in Table 2.

3.2 Distribution and correlation

The distribution of negatively framed test items was similar to that of the positively framed twinned items, however, four test items (three of which were the modifications for the item-'*doing things one wants to do*') out of nine had a response distribution of <5% at the most severe level. Table 3 displays the response distribution across the three positively framed items and their modified test items. Test items showed moderate correlations with their positive counterparts (0.40 to 0.55), but stronger correlations (>0.70) with other EQ-HWB items, such as lonely, unsupported, sad/depressed, look forward and cope. Table 4 displays the full correlation matrix between the nine test items and the original 25 items of the EQ-HWB.

3.3 Test-retest reliability

Test-retest reliability was assessed for 282 respondents who reported no changes in health, social circumstances, or overall quality of life during the 14-day recall period. Weighted Cohen's kappa values presented in Table 5 were in the moderate range (Landis & Koch, 1977), with slightly higher values for the negatively framed test items (0.54 to 0.64) compared to the positively framed twinned items (0.45 to 0.51).

3.4 Known-group differences

Known-group differences summarized in Table 6 revealed slightly better performance for the negatively framed test items, with effect sizes ranging from 0.47 to 0.73 for the VAS score

group (VAS ≥ 80), compared to 0.45 to 0.62 for the positively framed items. In the comorbidities group, the effect size for the negatively framed items ranged from 0.24 to 0.54, whereas the effect size for the positively framed items was between 0.17 and 0.26. Results of the Wilcoxon rank-sum test, presented in Table 7, showed significant differences across both groups for all items.

3.5 Confirmatory factor analysis

CFA results summarized in Table 8, showed excellent model fit (RMSEA: 0.03 to 0.05; TLI and CFI > 0.99).

3.6 Effect of notification on response pattern

Chi-square test results indicated a statistically significant association between notification status and response patterns for the item '*Did you feel accepted by others?*' ($\chi^2 = 15.2$, $p = 0.004$), which appeared first in the sequence after the notification. However, no statistically significant associations were observed for the other two positively framed items, '*Did you feel good about yourself?*' ($\chi^2 = 5.6$, $p = 0.229$) and '*Could you do the things you wanted to do?*' ($\chi^2 = 8.6$, $p = 0.071$)."

4. Discussion

The aims of this study were to first evaluate negatively framed versions of the three positively framed items, and second to assess the impact of flagging notifications on respondents' interpretation of these items. The findings suggest that, overall, the negatively framed items performed equally to original positively framed EQ-HWB items in several key psychometric evaluations. The distribution of responses across the items revealed no significant issues with skewness or monotonicity, reinforcing the stability and comprehensibility of the measure. The moderate correlation values observed within and across dimensions indicated that the items were sufficiently related within their respective subdomains, while no significant issues of cross-dimensional overlap were found. Notably, test-retest reliability was slightly higher for the negatively framed test items, suggesting that these items may provide clearer and more stable responses over time, possibly due to clearer interpretation and reduced response bias when framed negatively since this is consistent with the other EQ-HWB items. The known-group differences observed for the negatively framed test items also supported this further indicating a slightly better sensitivity of the negatively framed items to variations in health

status. While the groupings used offer a pragmatic approach to exploring differences, they may not fully capture the broader context of domains such as belonging, self-worth, and meaningful activities. These domains extend beyond core health measures and delve into aspects of social-care and carer-related QoL, which are influenced by factors outside traditional health parameters. As a result, judging their performance solely on their ability to distinguish between groups defined by health status or the number of reported health conditions may be overly restrictive. The overall results were consistent with other studies where negative framing was shown to improve item sensitivity and reduce potential confusion due to inconsistent interpretations of positively framed questions (Brazier et al., 2020).

The results of the CFA further confirm that the factor structure of the EQ-HWB remained robust after the modification, with excellent model fit indices. This suggests that the negative framing did not introduce construct irrelevance or disrupt the instrument's ability to measure the intended dimensions of health and well-being. The notification strategy, which flagged the transition to positively framed items, was found to have a significant impact on the first positively framed item, but not for the remaining two items. This suggests that the effect of the notification may be item-specific, highlighting the potential benefits of priming respondents for specific items but also suggesting that additional strategies may be needed to ensure consistent understanding across all items.

This is ongoing work, and further analyses will be conducted to extend the findings, and contribute to a more comprehensive understanding of item framing effects and their implications for measurement validity. Given that the samples testing Aims 1 and 2 differ, variation in demographics may lead to differences in response distributions. Future analyses will examine whether these differences persist after controlling for demographic factors such as age, gender, and health status. Additionally, identifying potential response inconsistencies by classifying responses as ‘potentially an error’ based on unlikely patterns relative to other known respondent data will be explored. Cross-tabulations of key items may help in detecting such inconsistencies and ensuring response validity.

Another area for further work is evaluating the justification for transitioning from positive to negative item framing in the EQ-HWB. This includes assessing whether notifying respondents

of framing changes is sufficient to mitigate response inconsistencies or if a full transition to a consistent framing approach is warranted. Additionally, as the current known group comparisons rely on health-based classifications, alternative subgroupings that better capture broader well-being dimensions such as belonging, self-worth, and meaningful activities will be considered.

An important limitation to this data collection is that it was opportunistically added to an existing data collection with other primary aims. Therefore the negatively framed alternative items were added at the end of the other EQ-HWB items, which is not where they would appear in the instrument if the items were replaced. If item changes were recommended, the characteristics of the chosen items could be tested as part of the 25-item system. The overall survey also included a range of other outcome measures including the EQ-5D-5L and ASCOT-SCT4 before the completion of the EQ-HWB, which could lead to order effects.

5. Conclusion

The overall findings support the idea that item framing is an important factor in survey design, a point well-documented in the extensive literature on framing effects in survey design. Modifying items to reduce potential biases, as highlighted in this study, could enhance the validity and interpretability of QoL measures like the EQ-HWB. The measurement properties of the negatively framed items were comparable, and in some cases better, than the original items. This indicates their potential to enhance measurement consistency in chronic condition populations. The evidence from this quantitative investigation can be used alongside other international mixed methods studies to inform whether the modified items should replace the positive items currently included in the EQ-HWB

6. Suggested ECR meeting discussion points

a. What types of evidence, and how much evidence, do you think is required to make a switch between positively and negatively framed items?
b. What other analyses could be included in guidance around testing positively and negatively worded items?

7. References

Brazier, J. E., Mulhern, B. J., Bjorner, J. B., Gandek, B., Rowen, D., Alonso, J., ... & Ware, J. E. (2020). Developing a new version of the SF-6D health state classification system from the SF-36v2: SF-6Dv2. *Medical care*, *58*(6), 557-565.

Brazier, J., & Tsuchiya, A. (2015). Improving cross-sector comparisons: going beyond the health-related QALY. *Applied health economics and health policy*, 13, 557-565.

Brazier, J., Peasgood, T., Mukuria, C., Marten, O., Kreimeier, S., Luo, N., ... & Rejon-Parrilla, J. C. (2022). The EQ-HWB: overview of the development of a measure of health and wellbeing and key results. *Value in Health*, *25*(4), 482-491.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, *70*(4), 213.

Cohen, J. (1992). A power primer. Psycholigcal Bulletin, 112 (1), 155–159.

Ebesutani, C., Drescher, C. F., Reise, S. P., Heiden, L., Hight, T. L., Damon, J. D., & Young, J. (2012). The Loneliness Questionnaire–short version: An evaluation of reverse-worded and non-reverse-worded items via item response theory. *Journal of Personality Assessment*, *94*(4), 427-437.

Harris, T., & Hardin, J. W. (2013). Exact Wilcoxon signed-rank and Wilcoxon Mann–Whitney ranksum tests. *The Stata Journal*, *13*(2), 337-343.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1-55.

Kuharić, M., Pickard, A. S., Mukuria, C., & Finch, A. P. (2024). The Measurement Properties of the EQ Health and Well-being and EQ Health and Well-Being Short Form in Italian Population: A Comparative Study With EQ-5D-5L. *Value in Health*.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374.

Masutti, S., Falivena, C., Purba, F. D., Jommi, C., Mukuria, C., & Finch, A. P. (2024). Content validity of the EQ-HWB and EQ-HWB-S in a sample of Italian patients, informal caregivers and members of the general public. *Journal of Patient-Reported Outcomes*, *8*(1), 36.

McDool, E., Mukuria, C., & Peasgood, T. (2024). Psychometric performance of the EQ Health and Wellbeing short (EQ-HWB-S) in a UK population sample. *Value in Health*.

Mukuria, C., Peasgood, T., McDool, E., Norman, R., Rowen, D., & Brazier, J. (2023). Valuing the EQ health and wellbeing short using time trade-off and a discrete choice experiment: a feasibility study. *Value in Health*, *26*(7), 1073-1084.

Peasgood, T., Mukuria, C., Brazier, J., Marten, O., Kreimeier, S., Luo, N., ... & Bjørner, J. (2022). Developing a new generic health and wellbeing measure: psychometric survey results for the EQ-HWB. *Value in Health*, *25*(4), 525-533.

Ponce, F. P., Irribarra, D. T., Vergés, A., & Arias, V. B. (2022). Wording effects in assessment: Missing the trees for the forest. *Multivariate Behavioral Research*, *57*(5), 718-734.

Salazar, M. S. (2015). The dilemma of combining positive and negative items in scales. *Psicothema*, *27*(2), 192-199.

Sauro, J., & Lewis, J. R. (2011, May). When designing usability questionnaires, does it hurt to be positive? In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2215-2224).

Sonderen, E. V., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PloS one*, *8*(7), e68967.

StataCorp. (2023). Stata Statistical Software: Release 18. College Station, TX: StataCorp LLC.

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of marketing research*, *45*(1), 116-131.

Table 1: Positively framed EQ-HWB questions with corresponding modified negatively framed test questions

In the last 7 days did you:	None of the time	Only occasionally	Some of the time / Sometimes	Often	Most or all of the time
Did you feel accepted by others? *(e.g. felt like you were able to be yourself and that you belonged)*					
feel excluded by others? (e.g. felt like you did not belong or you were not able to be yourself)	☐	☐	☐	☐	☐
feel left out?	☐	☐	☐	☐	☐
feel isolated?	☐	☐	☐	☐	☐
Did you feel good about yourself?					
feel negative about yourself?	☐	☐	☐	☐	☐
feel bad about yourself?	☐	☐	☐	☐	☐
feel like a failure?	☐	☐	☐	☐	☐

In the last 7 days how much difficulty did you have:	**No difficulty**	**Slight difficulty**	**Some difficulty**	**A lot of difficulty**	**Unable**
Could you do the things you wanted to do?					
doing enjoyable activities? (e.g., leisure, hobbies)	☐	☐	☐	☐	☐
doing the things you wanted to do?	☐	☐	☐	☐	☐
doing things you found rewarding?	☐	☐	☐	☐	☐

Table 2: Socio-demographic characteristics

Socio-demographic characteristics		Baseline	(%)	Follow-up	(%)
Total sample size at baseline			812	495	61%
Age	18-24 years	66	8	24	4.9
	25-34 years	144	17.7	81	16.4
	35-44 years	133	16.4	82	16.6
	45-54 years	128	15.8	75	15.2
	55-64 years	134	16.5	85	17.2
	65-74 years	127	15.6	88	17.8
	75+ years	80	9.9	60	12.1
Gender	Female	425	52	251	50.7
	Male	386	48	244	49.3
	Prefer not to say	1	0	0	0
Country of birth	Australia	674	83	393	79.4
Education	Year 12 or below	229	28.2	131	26.5
	Certificate or Diploma	248	30.5	158	31.9
	Bachelor's degree or higher	335	41.3	206	41.6
Employment	Employed full-time	345	42.5	198	40
	Employed part-time	111	13.7	67	13.5
	Retired	222	27.3	155	31.3
	Other	134	16.5	75	15.2
Self-report health condition	Type 1 diabetes	51	6	26	5.3
	Type 2 diabetes	152	19	106	21.4
	Anxiety	81	10	53	10.7
	Depression	35	4	26	5.3
	Both anxiety and depression	88	11	59	11.9
	Chronic heart condition	204	25	112	22.6
	Asthma	171	21	90	18.2
	Chronic obstructive pulmonary disease (COPD)	30	4	23	4.6
	Comorbidity reported other than main condition	456	57	293	59.2

Table 3: Response distribution for the positively framed EQ-HWB items with corresponding modified negatively framed test items

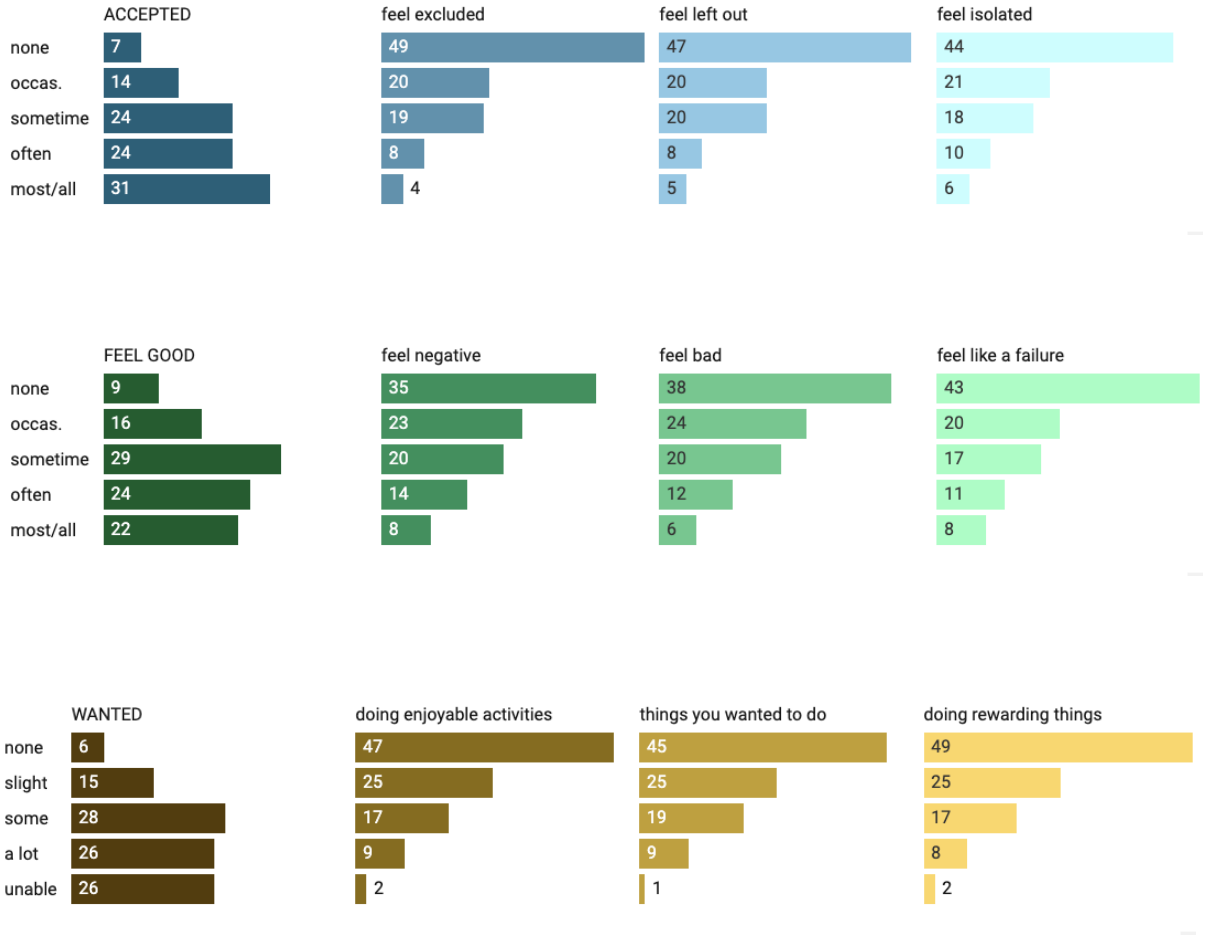
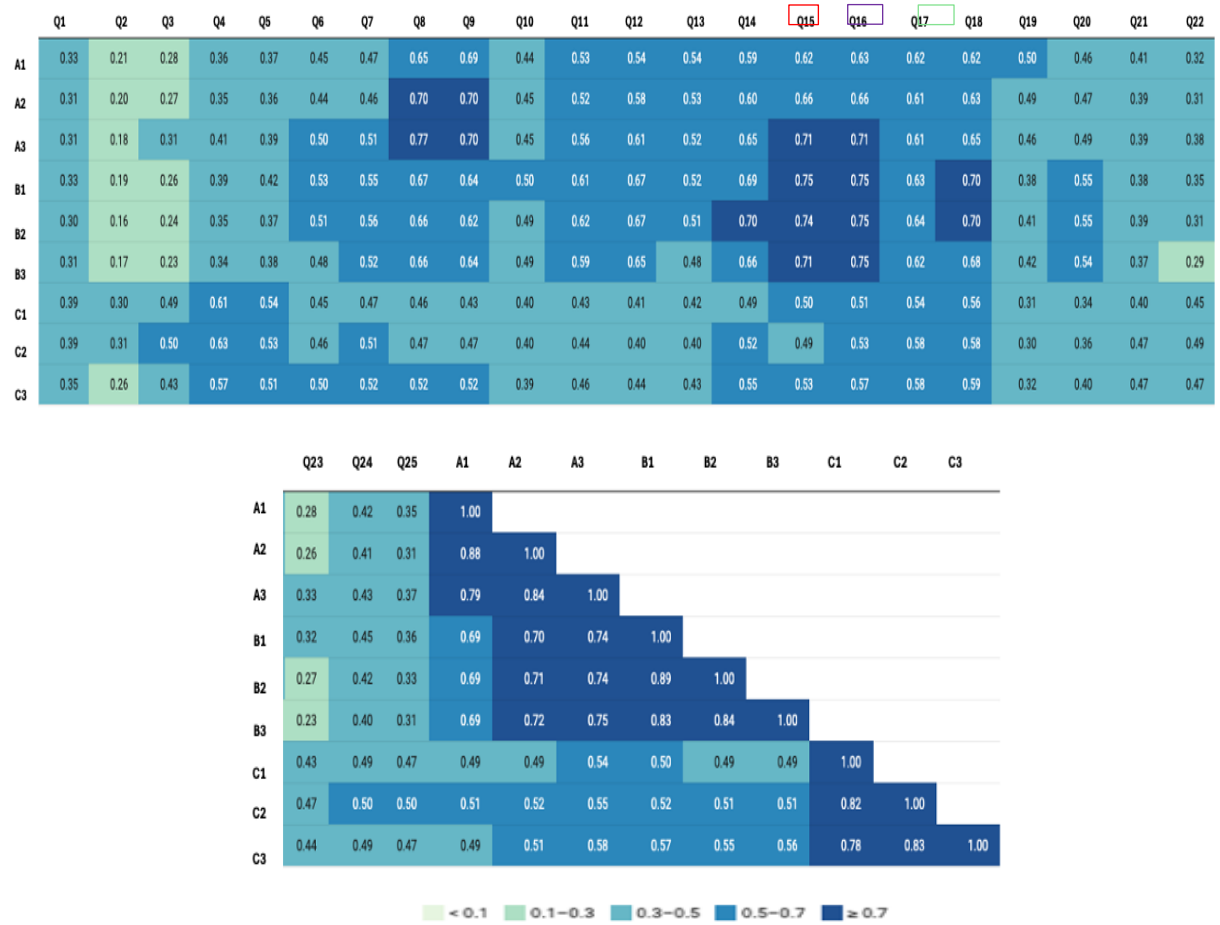


Table 4: Correlation matrix between modified test items and EQ-HWB



EQ-HWB 25 items=Q1: See; Q2: Hear; Q3: getting around; Q4: Daily activity; Q5: Personal care; Q6: Sleep; Q7: Exhausted; Q8: Lonely; Q9: Unsupported; Q10: Remembering; Q11: Concentrating; Q12: Anxious; Q13: Unsafe; Q14: Frustrated; Q15:Depressed; Q16: Look forward; Q17: Control; Q18: Cope; Q19: Accepted; Q20: Feel good; Q21: Do things wanted to do; Q22: Pain (frequency); Q23: Pain (severity); Q24: Discomfort (frequency); Q25: Discomfort (severity).
Test items= A1: Feel excluded; A2: Feel left out; A3: Feel isolated; B1: Feel negative; B2: Feel bad; B3: Feel like a failure; C1: Doing enjoyable activities; C2: Doing things you wanted to do; C3: Doing rewarding things.

Table 5: Test-retest reliability results using Weighted Cohen's kappa

EQ-HWB items	Agreement	Weighted kappa
ACCEPTED	0.81	0.45
feel excluded	0.87	0.54
feel left out	0.88	0.60
feel isolated	0.87	0.60
FEEL GOOD	0.83	0.51
feel negative	0.87	0.62
feel bad	0.88	0.64
feel like a failure	0.87	0.64
WANTED	0.83	0.47
doing enjoyable activities	0.87	0.53
things you wanted to do	0.87	0.55
doing rewarding things	0.88	0.54

N=282; <0.00 (poor); 0.00-0.20 (slight); 0.21-0.40 (fair); 0.41-0.60 (moderate); 0.61-0.80 (substantial); 0.81-1.00 (almost perfect) (Landis & Koch, 1977)

Table 6: Known group differences

EQ-HWB items	Cohen's D	
	By VAS score	By comorbidity
ACCEPTED	0.45	-0.17
feel excluded	-0.48	0.30
feel left out	-0.47	0.28
feel isolated	-0.61	0.38
FEEL GOOD	0.59	-0.2
feel negative	-0.66	0.32
feel bad	-0.59	0.24
feel like a failure	-0.62	0.28
WANTED	0.62	-0.26
doing enjoyable activities	-0.63	0.44
things you wanted to do	-0.73	0.54
doing rewarding things	-0.66	0.49

Effect sizes (Cohen's D) between 0.2-0.49 are considered small, 0.5-0.79 moderate, and ≥0.8 large (Cohen 1992)

Table 7: Known group differences results using Wilcoxon rank-sum test

EQ-HWB items	Groups	Group 1	Group 2	Z-score	Exact p-value
ACCEPTED	VAS score	214	598	6.811	<0.0001
	Comorbidity	456	356	-2.632	0.0084
feel excluded	VAS score	214	598	-7.409	<0.0001
	Comorbidity	456	356	4.275	<0.0001
feel left out	VAS score	214	598	-7.239	<0.0001
	Comorbidity	456	356	3.821	0.0001
feel isolated	VAS score	214	598	-8.756	<0.0001
	Comorbidity	456	356	4.954	<0.0001
FEEL GOOD	VAS score	214	598	8.510	<0.0001
	Comorbidity	456	356	-3.047	0.0023
feel negative	VAS score	214	598	-9.348	<0.0001
	Comorbidity	456	356	4.196	<0.0001
feel bad	VAS score	214	598	-8.441	<0.0001
	Comorbidity	456	356	3.061	0.0022
feel like a failure	VAS score	214	598	-8.837	<0.0001
	Comorbidity	456	356	3.358	0.0008
WANTED	VAS score	214	598	9.069	<0.0001
	Comorbidity	456	356	-3.972	0.0001
doing enjoyable activities	VAS score	214	598	-9.804	<0.0001
	Comorbidity	456	356	6.550	<0.0001
things you wanted to do	VAS score	214	598	-10.483	<0.0001
	Comorbidity	456	356	7.815	0.0001
doing rewarding things	VAS score	214	598	-9.997	<0.0001
	Comorbidity	456	356	6.966	<0.0001

Table 8: Confirmatory factor analysis

Observed Variables	Factor Loading	Standard Error (SE)	z-value	p-value	Residual Variance
ACCEPTED	1 (fixed)	-	-	-	1.228
feel excluded	-1.787	0.126	-14.16	<0.001	0.253
feel left out	-1.974	0.138	-14.29	<0.001	0.098
feel isolated	-1.861	0.134	-13.92	<0.001	0.42

Model fit indices: RMSEA = 0.034; TLI = 0.998; CFI = 0.999

Observed Variables	Factor Loading	Standard Error (SE)	z-value	p-value	Residual Variance
FEEL GOOD	1 (fixed)	-	-	-	1.026
feel negative	-1.75	0.095	-18.45	<0.001	0.191
feel bad	-1.721	0.093	-18.55	<0.001	0.136
feel like a failure	-1.694	0.094	-17.97	<0.001	0.366

Model fit indices: RMSEA = 0.032; TLI = 0.998; CFI = 0.999

Observed Variables	Factor Loading	Standard Error (SE)	z-value	p-value	Residual Variance
WANTED	1 (fixed)	-	-	-	1.117
doing enjoyable activities	-1.69	0.125	-13.53	<0.001	0.283
things you wanted to do	-1.8	0.131	-13.78	<0.001	0.163
doing rewarding things	-1.701	0.125	-13.63	<0.001	0.272

Model fit indices: RMSEA = 0.052; TLI = 0.993; CFI = 0.998

RMSEA (root mean square error of approximation) <0.6 taken as good, CFI (comparative fit index)>0.95 taken as good, and TLI (Tucker Lewis index)>0.95 taken as good. (Hu & Bentler, 1999).