Examining the Psychometric Properties of Double-Barreled Items in the EQ-HWB/EQ-HWB-S among Caregivers and Care Recipients

Maja Kuharic[1,2], Brendan Mulhern[3], A. Simon Pickard[2]

1 Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, 625 N Michigan Ave, Chicago, Illinois, USA.

2 Department of Pharmacy Systems, Outcomes and Policy. College of Pharmacy, University of Illinois Chicago. 833 South Wood Street (MC 871). Chicago, Illinois , USA

3 Centre for Health Economics Research and Evaluation, University of Technology Sydney, Sydney, NSW, Australia

Abstract

Objectives: This study aimed to assess the psychometric properties of two double-barreled items, "Concentrating/Thinking Clearly" and "Walking Inside/Outside", in the EQ-HWB/EQ-HWB-S, compared to their single-domain counterparts from the E-QALY project, among caregivers and patients. These double-barreled items were created by merging four single-domain items from the initial E-QALY project; however, their measurement properties have not been previously evaluated.

Methods: A secondary analysis of cross-sectional data was conducted on 504 caregiver-patient dyads in the US using an online panel between August 2022 and February 2023. Participants completed the EQ-HWB/EQ-HWB-S, and the four single-domain items from the E-QALY project. Separately in a sample of caregivers and patients, we conducted the psychometric analysis between the double-barreled items and their single-domain counterparts: 1) floor/ceiling effects 2) correlations using Spearman's rank correlation; 3) level of agreement; 4) confirmatory factor analysis; 5) item response theory-based models and 6) differential item functioning.

Results: Double-barreled items showed shifted response distributions, lower ceiling effects, strong correlations (r = 0.70-0.78), and substantial agreement (κ = 0.69-0.79) with their single-domain counterparts. CFA demonstrated strong factor loadings (0.877-0.976) for all items. The IRT analysis revealed that the double-barreled items provided comparable levels of information to the single-domain items across the latent trait range, with strong discrimination parameters (a=3.02-3.62). DIF analysis showed negligible overall DIF for most item pairs.

Conclusion: The double-barreled items "Concentrating/Thinking Clearly" and "Walking Inside/Outside" in the EQ-HWB/EQ-HWB-S demonstrated strong psychometric properties, including convergent validity, structural validity, and precision in measuring the underlying constructs, while generally functioning similarly to their single-domain counterparts. The use of carefully constructed double-barreled items in may provide a more efficient approach to capturing related health domains without compromising measurement precision.

Introduction

The measurement of health-related quality of life (HRQoL) has been a cornerstone of health outcomes research for decades, with the EuroQol Group's EQ-5D instrument being widely adopted in health economic evaluations, clinical research, and population health studies.[1] [2] As the field advanced, there has been growing recognition of the need to capture a more comprehensive range of health and well-being domains.[3] This recognition extends beyond traditional health outcomes to include aspects of social care and the impact on caregivers. [4] In response to this need, the "Extending the QALY" (E-QALY) project was initiated with support from the EuroQol Group, which led to the development of the EQ Health and Wellbeing (EQ-HWB) and its short form (EQ-HWB-S).[5-7] These new instruments were designed to capture a wider range of dimensions relevant to health, social care, and carer-related quality of life, while still maintaining practicality for use in economic evaluation and other applications across health and social care sectors. The instruments were developed through an international collaboration involving six countries (Argentina, Australia, China, Germany, United Kingdom, and the United States) and have undergone extensive face validation and psychometric testing in other countries since then.[5-7]

During the initial development phase of the EQ-HWB, the E-QALY project included several individual items that were later merged into double-barreled items in the current versions of the EQ-HWB and EQ-HWB-S.[5-7] Double-barreled items, also referred to as composite, compound or multidimensional items, are items that address more than one issue or concept while only allowing for a single response. [8] For instance in EQ-HWB, separate items assessing concentration and clarity of thinking were combined into a single double-barreled item. Similarly, items evaluating mobility inside and outside the home were merged. While this approach can reduce instrument’s length and respondent burden, it raises important questions about the psychometric properties of these merged items.

The use of double-barreled items in health measurement scales has been a subject of ongoing debate in psychometric research. An argument in favor of the use of double-

barreled items would be in situations where the components of a double-barreled item are highly correlated or conceptually related.[9] Some studies have found that carefully constructed double-barreled items can effectively capture complex health concepts without compromising measurement precision. [10] For example, quality of life measures often include double-barreled items that combine closely related physical and emotional functioning aspects, as these domains are frequently intertwined in the lived experience of respondents. [11] However, for some constructs, these items can present significant challenges in terms of validity and reliability, potentially introducing ambiguity and confusion for respondents.[9] When confronted with a double-barreled item, respondents might have different opinions or experiences related to each component of the question, making it difficult to provide a single response to multiple concepts. [12] This can lead to increased measurement error, as respondents may interpret the item differently or provide responses that do not accurately reflect their true status. [13] For example, a study by Engel et al. found that the EQ-5D-5L pain/discomfort dimension captures aspects of pain more than aspects of discomfort, possibly due to the absence of descriptors or because pain is mentioned first in the composite item.[14] The key lies in carefully designing and psychometric testing these items to ensure they maintain measurement validity and reliability.

Despite the inclusion of double-barreled items in the current versions of the EQ-HWB and EQ-HWB-S, their psychometric properties have not been directly compared to their single-domain counterparts from the E-QALY project. This is an essential step in establishing the validity and reliability of the merged items, as the psychometric properties of double-barreled items can differ from those of their single-domain components.[12, 13, 15] In the context of health utility measurement, where precise quantification of health states is essential for economic evaluations and decision-making, it is important to understand the measurement characteristics of double barreled items to inform the health state descriptions used in valuation. Our study aims to address this gap by assessing the psychometric properties of two double-barreled items in the EQ-HWB/EQ-HWB-S: "Concentrating/Thinking Clearly" and "Walking Inside/Outside". We will compare these to

their single-domain counterparts from the E-QALY project, examining their psychometric performance among both patients and caregivers. Our findings provide valuable insights into the performance of double-barreled items in health utility measurement and inform the future development and use of these important instruments.

Methods

Study Design and Participants

This study is a secondary analysis of cross-sectional data collected from 504 patient-caregiver dyads in the United States between August 2022 and February 2023. The study was approved by the Institutional Review Board at the University of Illinois Chicago (#2022-0490). Participants were recruited through an online panel using quota sampling to ensure diversity in race, age, and gender, reflecting the demographics of informal caregivers in the U.S. [16]

Eligible participants included caregivers aged 18 years or older who had provided unpaid care to a relative or friend aged 18 years or older for at least one hour per week over the past six months. Patients/care recipients were required to confirm that they had received care from their caregiver within the previous six months and were at least 18 years old. Informed consent was obtained from all participants.

Measures

Both caregivers and patients completed the EQ-HWB. [2,3] Immediately following, participants responded to four individual items from the original E-QALY project. These items corresponded to two specific double-barreled items in the current EQ-HWB/EQ-HWB-S:

1. Cognition (Response levels: None of the time / Only occasionally / Sometimes / Often / Most or all of the time):
    - Double-barreled: "*Did you have trouble concentrating or thinking clearly*"

- Single-domain:
    1. "*I found it hard to concentrate*"
    2. "*I had trouble thinking clearly"*

2. Mobility (Response levels: No difficulty / Slight difficulty / Moderate difficulty / A lot of difficulty / Unable to do):
    - Double-barreled: "*How much difficulty did you have getting around inside and outside?"*
    - Single-domain:
        1. *" How well were you able to get around inside your home?"*
        2. *" How well were you able to get around outside?"*

Demographic and clinical characteristics of the participants were also collected, including age, gender, race/ethnicity, education level, and the presence of chronic conditions.

Data Collection and Quality

The survey was developed and administered using the Qualtrics platform (Provo, UT, USA). To minimize order effects, all measures in the survey were presented in a randomized sequence. [17] The sequential linking method was used for data collection, allowing caregivers and patients to complete the survey consecutively during a single session, resulting in more efficient data collection. [18]

Several measures were implemented to ensure data quality and validity. Validity checks based on demographic and relationship variables were used to confirm the authenticity of caregiver-patient dyads and prevent situations where one participant completed the survey for both members of the dyad. [18] [19] Attention-check questions and response time monitoring were employed to identify and exclude inattentive respondents. [20] For survey bots, we implemented of Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA), cookies, I.P. address, geolocation

data before data collection, invisible 'honeypot' questions, misspelled words and image-based text questions. [19-22]

Data Analysis

We conducted separate analyses for caregivers and patients to account for potential differences in item functioning between these groups. Descriptive statistics summarized participant characteristics and item response distributions.

Response Distributions and Ceiling/Foor Effect

We calculated the percentage of respondents selecting each response option for both single-domain and double-barreled items. Floor effects were defined as the percentage selecting the worst possible response, while ceiling effects were the percentage choosing the best possible response.[23] We considered floor or ceiling effects to be present if more than 35% of respondents achieve the lowest or highest possible score, respectively.[23] To compare the endorsement of double-barreled items with single-domain items, we calculated the difference in the percentage of respondents selecting each response option between the double-barreled item and the average of its corresponding single-domain items.[24]

Correlations and Agreement

Spearman rank-order correlations were calculated to assess the associations between the double-barreled items and their single-domain counterparts. Correlations were interpreted as weak (0.10-0.29), moderate (0.30-0.49), or strong (≥0.50). [25] Weighted kappa coefficients were computed to evaluate the level of agreement between the items. [26, 27] Kappa values were interpreted as slight (0.01-0.20), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80), or almost perfect (0.81-1.00). [28]

Confirmatory Factor Analysis (CFA)

CFA was performed to examine the structural validity of the double-barreled items in comparison to their single-domain counterparts.[29] Two separate two-factor models (Cognition and Mobility) were specified for caregivers and patients. Given the ordinal nature of responses, we used the weighted least squares means and variance adjusted (WLSMV) estimator with theta parameterization.[29] While we report standard model fit indices (comparative fit index (CFI), Tucker-Lewis index (TLI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR) [30], our primary focus was on the standardized factor loadings.[31]

Item Response Theory (IRT)

Graded response models (GRMs) were constructed to evaluate the item characteristics and information functions.[32] We examined item characteristic curves (ICCs) and item information functions to compare discrimination and difficulty parameters across the latent trait continuum.[33] ICCs display the probability of endorsing each response category as a function of the latent trait, while item information functions indicate the precision of the items in measuring the latent trait at different levels. Discrimination parameters >0.8 were considered acceptable, with values >1.3 indicating highly discriminating items.[34]

Differential Item Functioning (DIF)

We conducted a novel application of DIF analysis to compare how respondents answer single-domain items versus their corresponding double-barreled items. This approach compared the probability of selecting different response options for a single-domain item (referent) versus the corresponding double-barreled item (focal) while controlling for the respondent's overall trait level (e.g., mobility or cognitive functioning), as estimated by the IRT model. [35] Using a hybrid ordinal logistic regression method which combines IRT and logistic regression to detect both uniform and non-uniform DIF, we compared three nested models: [36]

- Model 1: Item response predicted by trait level only

- Model 2: Item response predicted by trait level and item type (single-domain vs. double-barreled)
- Model 3: Item response predicted by trait level, item type, and their interaction

This method allowed us to detect both uniform DIF (which occurs when the difference in item functioning between the referent and focal items is consistent across all levels of the trait) and non-uniform DIF (the difference in item functioning varies across trait levels). [11] We quantified DIF magnitude using changes in McFadden's pseudo-R^2 between models, with $\Delta R^2 \geq 0.02$ indicating meaningful DIF. [37, 38]

Statistical analyses were performed using SAS Version 9.4 for descriptive statistics, correlations, level of agreement and IRT; Mplus Version 8.4 for CFA; and DIF analysis with using the lordif R package. [39]

Results

Participant Characteristics

From an initial pool of 4,714 survey participants who started the survey, screening process excluded 2,651 participants (56.2%) during eligibility screening, 957 (20.3%) following validity checks, 317 (6.7%) due to quality checks, and 285 (6.0%) to meet race and gender quotas. Ultimately, 504 care recipient-caregiver dyads (10.7% of initial participants) successfully completed the survey and met all inclusion criteria (**Table 1**).

Caregivers were predominantly female (57.5%), with a mean age of 49.2 years (SD = 15.4). The majority were White (73.2%), employed (61.7%), and married or living with a partner (69.4%). This demographic profile closely aligns with that of informal caregivers in the United States, where approximately 58% of caregivers are women. [40] Care recipients were older (mean age 62.7 years, SD = 18.9), with a relatively even gender distribution (52.4% female). Most care recipients were also White (71.8%), retired or homemakers (46.2%), and married or living with a partner (49.2%). Caregivers reported a mean EQ-5D-5L Index score of 0.73 (SD = 0.28) and EQ-HWB-S Index score of 0.67 (SD = 0.26). Care recipients showed

lower scores (EQ-5D-5L: 0.43, SD = 0.40; EQ-HWB-S: 0.47, SD = 0.03), reflecting poorer health status.

The majority of caregivers (87.1%) identified as primary caregivers, with 34.5% caring for a spouse/partner and 29.8% caring for a parent. Most (68.3%) lived in the same household as the care recipient. Caregiving intensity was high, with 69.9% categorized as Level 4 or 5 on the Level of Care Index. The most common reasons for providing care were long-term physical conditions (58.9%) and old age/aging (47.0%).

Descriptive Statistics and Response Distributions

Double-barreled items generally showed lower percentages in extreme response categories and higher percentages in middle response options compared to single-domain items (**Table 2**). For the Concentrating/Thinking Clearly items, the double-barreled version showed increases of 3.77% and 3.67% in the "Sometimes" and "Often" categories for caregivers, and 6.95% in the "Often" category for patients. The Walking Inside/Outside double-barreled item showed smaller differences, with the largest increase of 2.78% in the "A lot of difficulty" category for patients.

Ceiling effects were generally lower for double-barreled items. For caregivers, a ceiling effect was observed only in the "Thinking Clearly" single item (41.87% selecting "None of the time"), while the double-barreled item (33.33%) fell below the threshold. For patients, no ceiling effects were observed in any Concentrating/Thinking Clearly items. For the Walking Inside/Outside items, ceiling effects were present in all caregiver responses (Walking Inside: 62.90%, Walking Outside: 56.75%, Walking Inside/Outside: 60.12% selecting "No Difficulty"). For patients, neither floor nor ceiling effects were observed in any Walking Inside/Outside items.

Correlations and Level of Agreement

Spearman rank-order correlations between the double-barreled items and their single-domain counterparts were strong and statistically significant ($p < 0.001$) in both

caregiver and patient samples (**Table 3**). For the Concentrating/Thinking Clearly item group, correlations ranged from 0.74 to 0.78, with the double-barreled item showing strong correlations with both single-domain items in caregivers (r = 0.78 and 0.76) and patients (r = 0.74 for both). For the Walking Inside/Outside item group, correlations ranged from 0.70 to 0.75, with the double-barreled item correlating strongly with both single-domain items in caregivers (r = 0.70 and 0.75) and patients (r = 0.75 and 0.71).

Weighted kappa coefficients indicated substantial to almost perfect agreement between the double-barreled items and their single-domain counterparts (**Table 3**). For the Concentrating/Thinking Clearly item group, kappa values ranged from 0.72 to 0.79, while for the Walking Inside/Outside item group, they ranged from 0.69 to 0.76.

Confirmatory Factor Analysis

The CFA results showed strong factor loadings for all items in both caregiver and patient samples (see **Table 4**). For caregivers, factor loadings ranged from 0.877 to 0.974, with the double-barreled items (Concentrating/Thinking Clearly and Walking Inside/Outside) showing slightly lower but still very high loadings (0.880 and 0.877, respectively) compared to their single-domain counterparts. The correlation between the Cognition and Mobility factors was moderate (r = 0.552) in the caregiver sample.

For patients, factor loadings ranged from 0.904 to 0.976, with the double-barreled items demonstrating high loadings (0.913 for Concentrating/Thinking Clearly and 0.948 for Walking Inside/Outside). Interestingly, the double-barreled Walking Inside/Outside item had a higher loading than one of its single-domain counterparts in the patient sample. The correlation between the Cognition and Mobility factors was weaker (r = 0.375) in the patient sample compared to the caregiver sample.

Item Response Theory (IRT)

The ICCs for the double-barreled items and their single-domain counterparts exhibited similar shapes and locations along the latent trait continuum (**Figures 1-2**). For the

Concentrating/Thinking Clearly item group, the ICCs were close to each other for both caregivers and patients. For the Walking Inside/Outside item group, the ICCs for the single-domain items were almost overlapping, and the double-barreled item's ICC was similar, with slight differences in the probabilities of endorsing middle response options.

IIT revealed that the double-barreled items and their single-domain counterparts provided similar levels of information across the latent trait range (**Figures 1-2**).

Discrimination parameters for all items were above 0.8, indicating acceptable discrimination (**Table 5**). The Concentrating/Thinking Clearly items showed high discrimination (range: 3.02-6.60) across both caregivers and patients, with the double-barreled item demonstrating strong discrimination in caregivers (a = 3.62) and patients (a = 3.02). Similarly, the Walking Inside/Outside items exhibited high discrimination (range: 3.11-8.63), with the double-barreled item showing strong discrimination in caregivers (a = 3.14) and patients (a = 3.11).

Differential Item Functioning (DIF)

For the Concentrating/Thinking Clearly item group, the overall DIF (ΔR^2) was negligible for both caregivers (0.0007 to 0.0047) and patients (0.0056 to 0.0079), indicating no meaningful difference in response probabilities between the single-domain and double-barreled items. Uniform and non-uniform DIF were also negligible ($\Delta R^2 < 0.02$) for both caregivers and patients, suggesting consistent item functioning across trait levels.

For the Walking Inside/Outside item group, the overall DIF was negligible for caregivers ($\Delta R^2 = 0.0003$ to 0.0013) and patients ($\Delta R^2 = 0$ to 0.0309). However, patients showed a small but notable overall DIF for the Walking Inside/Walking Outside-Inside pair ($\Delta R^2 = 0.0309$), with a small uniform DIF ($\Delta R^2 = 0.0241$), suggesting a slight difference in response probabilities and item functioning for this item pair.

Discussion

This study evaluated the psychometric properties of two double-barreled items in the EQ-HWB/EQ-HWB-S, "Concentrating/Thinking Clearly" and "Walking Inside/Outside", in comparison to their single-domain counterparts from the E-QALY project. Our findings provide strong evidence supporting the validity and reliability of these items in both caregiver and patient samples, suggesting that carefully constructed double-barreled items can effectively capture related health domains without compromising measurement precision.

Response distributions revealed that double-barreled items slightly favored middle categories over extreme ones compared to single-domain items, particularly in the Concentrating/Thinking Clearly domain. This suggests a potential for capturing a wider range of the latent trait, which is crucial measurement property to inform the development of health utility measures requiring a wide range of health states for valuation purposes. However, the differences were generally small, indicating that double-barreled items largely maintain the response pattern of their single-domain counterparts while potentially offering enhanced sensitivity in the mid-range of health states. The strong correlations and substantial agreement between the double-barreled items and their single-domain counterparts provide compelling evidence of convergent validity. These findings, along with the strong factor loadings demonstrated in the CFA, support the structural validity of the double-barreled items and their ability to measure the intended latent constructs (Cognition and Mobility) in both caregiver and patient samples. This suggests that the double-barreled items can be used interchangeably with their single-domain counterparts without significant loss of information. The IRT analyses provide additional support for the psychometric strength of the double-barreled items. The comparable item characteristic curves and high discrimination parameters indicate that these items perform similarly to their single-domain counterparts in measuring the underlying constructs. It's worth noting that while all slope parameters were high, indicating good discrimination, the combined item slopes were relatively lower than their single counterparts within each domain. This subtle difference suggests that while the double-barreled items perform well, they may have

slightly less precision in discriminating between different levels of the latent trait compared to the single items. The minimal DIF observed for most item pairs in the DIF analysis indicates that the double-barreled items generally function similarly to their single-domain counterparts across different trait levels. While the small but notable DIF observed for the Walking Inside/Walking Outside-Inside pair in patients warrants further investigation, the overall results support the comparability of the double-barreled items to their single-domain counterparts in terms of item functioning and response probabilities.

These findings have important implications for the ongoing development of the EQ-HWB/EQ-HWB-S. The strong psychometric properties of the double-barreled items support their inclusion in the instruments, potentially allowing for a more efficient assessment of health and wellbeing without compromising measurement quality. This is particularly valuable in the context of the EQ-HWB-S, where brevity and comprehensiveness must be carefully balanced. From a broader perspective, this study contributes to the ongoing debate about the use of double-barreled items in health status measures. Traditionally, survey methodologists have cautioned against their use, arguing that they can introduce construct-irrelevant variance and potentially confuse respondents. [41] Our results suggest that carefully constructed double-barreled items can perform well psychometrically while offering the advantage of increased efficiency. This is particularly valuable in the context of patient-reported outcome measures, where reducing respondent burden while maintaining comprehensive assessment is a key consideration. From a practical perspective, the use of double-barreled items in the EQ-HWB/EQ-HWB-S aligns with a growing trend in patient-reported outcome measures to balance comprehensiveness with respondent burden. However, it is crucial to note that the effectiveness of double-barreled items likely depends on how closely related the combined concepts are and how they are perceived by respondents. It's important to distinguish between the measurement properties of these items and their potential use in valuation studies. While our results support the psychometric validity of the double-barreled items, their implications for health state valuation require separate consideration.

While our study has consistently referred to these items as "double-barreled," our findings prompt an interesting philosophical question about terminology. Given that our psychometric analyses support the combination of related concepts into a single item, we might reconsider the appropriateness of the term "double-barreled" in this context. As a point of reflection, we might ask: If psychometric evidence supports combining items, should we still consider them double-barreled? This consideration invites further debate on the nomenclature used in psychometric research and item development.

Several limitations of this study should be noted. First, our sample was limited to caregivers and patients in the United States, which may limit the generalizability of our findings to other cultural contexts. Cross-cultural validation studies would be valuable to ensure the performance of these items across diverse populations.[42] Second, while we employed a comprehensive set of psychometric analyses, additional methods such as longitudinal invariance testing could provide further insights into the stability of these items over time.[43] Third, qualitative research exploring how respondents interpret and respond to these double-barreled items could provide valuable complementary information to our quantitative findings. Finally, each factor was measured by only three items (two single-domain and one double-barreled). While this is the minimum number of items typically recommended for CFA, more items per factor could provide a more comprehensive assessment of the constructs and potentially reveal additional complexities in the factor structure. [31] The small number of items per factor may not capture the full breadth of the constructs. With three items per factor, the model is just identified, which limits our ability to assess overall model fit. Future research could benefit from incorporating the broader context of the EQ-HWB measure. Examining how these items perform within the wider dimensionality of the instrument could provide additional insights into their functioning and appropriateness.

In conclusion, this study provides strong evidence supporting the validity and reliability of the double-barreled items (Concentrating/Thinking Clearly and Walking Inside/Outside) in the EQ-HWB/EQ-HWB-S. As the EQ-HWB/EQ-HWB-S continue to be

refined and validated, these results contribute valuable insights to guide their development and use in diverse healthcare and research settings. The use of carefully constructed double-barreled items in may provide a more efficient approach to capturing related health domains without compromising measurement precision.

Acknowledgements: The EuroQol Research Foundation supported this research and original data collection through a dissertation research grant (Kuharic/Pickard).

References

1. Brooks, R. and E. Group, *EuroQol: the current state of play*. Health policy, 1996. **37**(1): p. 53-72.

2. Devlin, N.J. and R. Brooks, *EQ-5D and the EuroQol group: past, present and future*. Applied health economics and health policy, 2017. **15**: p. 127-137.

3. Brazier, J., et al., *Measuring and valuing health benefits for economic evaluation*. 2017: OXFORD university press.

4. Brazier, J.E., et al., *Future directions in valuing benefits for estimating QALYs: is time up for the EQ-5D?* Value in Health, 2019. **22**(1): p. 62-68.

5. Brazier, J., et al., *The EQ-HWB: Overview of the Development of a Measure of Health and Wellbeing and Key Results*. Value in Health, 2022. **25**(4): p. 482-491.

6. Carlton, J., et al., *Generation, Selection, and Face Validation of Items for a New Generic Measure of Quality of Life: The EQ-HWB*. Value in Health, 2022. **25**(4): p. 512-524.

7. Peasgood, T., et al., *Developing a New Generic Health and Wellbeing Measure: Psychometric Survey Results for the EQ-HWB*. Value in Health, 2022. **25**(4): p. 525-533.

8. Tourangeau, R., L.J. Rips, and K. Rasinski, *The psychology of survey response*. 2000.

9. Menold, N. and T. Raykov, *On the Relationship Between Item Stem Formulation and Criterion Validity of Multiple-Component Measuring Instruments*. Educ Psychol Meas, 2022. **82**(2): p. 356-375.

10. Böhnke, J.R. and T.J. Croudace, *Calibrating well-being, quality of life and common mental disorder items: psychometric epidemiology in public mental health research*. The British Journal of Psychiatry, 2016. **209**(2): p. 162-168.

11. Fayers, P.M. and D. Machin, *Quality of life: the assessment, analysis and reporting of patient-reported outcomes*. 2015: John Wiley & Sons.

12. Dillman, D.A., J.D. Smyth, and L.M. Christian, *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. 2014: John Wiley & Sons.

13. Yan, T. and R. Tourangeau, *Fast times and easy questions: The effects of age, experience and question complexity on web survey response times*. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 2008. **22**(1): p. 51-68.

14. Engel, L., et al., *What is measured by the composite, single-item pain/discomfort dimension of the EQ-5D-5L? An exploratory analysis*. Qual Life Res, 2023. **32**(4): p. 1175-1186.

15. Olson, K., *An examination of questionnaire evaluation by expert reviewers*. Field methods, 2010. **22**(4): p. 295-318.

16. Kuharic, M., et al., *Comparison of the EQ-HWB and EQ-HWB-S With Other Preference-Based Measures Among United States Informal Caregivers*. Value in Health, 2024. **27**(7): p. 967-977.

17. Rossi, P.H., J.D. Wright, and A.B. Anderson, *Handbook of survey research*. 2013: Academic press.

18. Ganong, L., L. Russell, and N. Stoddard, *Conducting Online Research With Dyads*. 2022: London.

19. Griffin, M., et al., *Ensuring survey research data integrity in the era of internet bots*. Qual Quant, 2022. **56**(4): p. 2841-2852.

20. Pei, W., et al. *Attention please: Your attention check questions in survey studies can be automatically answered*. in *Proceedings of The Web Conference 2020*. 2020.

21. Teitcher, J.E., et al., *Detecting, preventing, and responding to "fraudsters" in internet research: ethics and tradeoffs*. J Law Med Ethics, 2015. **43**(1): p. 116-33.

22. Storozuk, A., et al., *Got bots? Practical recommendations to protect online survey data from bot attacks*. The Quantitative Methods for Psychology, 2020. **16**(5): p. 472-481.

23. Terwee, C.B., et al., *Quality criteria were proposed for measurement properties of health status questionnaires*. Journal of clinical epidemiology, 2007. **60**(1): p. 34-42.

24. Menold, N., *Double barreled questions: An analysis of the similarity of elements and effects on measurement quality*. Journal of Official Statistics, 2020. **36**(4): p. 855-886.

25. Cohen, J., *Statistical power analysis for the behavioral sciences*. 2013: routledge.

26. Cicchetti, D.V., *Testing the normal approximation and minimal sample size requirements of weighted kappa when the number of categories is large*. Applied psychological measurement, 1981. **5**(1): p. 101-104.

27. Ranganathan, P., C.S. Pramesh, and R. Aggarwal, *Common pitfalls in statistical analysis: Measures of agreement*. Perspect Clin Res, 2017. **8**(4): p. 187-191.

28. Cohen, J., *A coefficient of agreement for nominal scales*. Educational and psychological measurement, 1960. **20**(1): p. 37-46.

29. Brown, T.A., *Confirmatory factor analysis for applied research*. 2015: Guilford publications.

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

31. Kline, R.B., *Principles and practice of structural equation modeling*. 2023: Guilford publications.

32. Samejima, F., *Graded response models*, in *Handbook of item response theory*. 2016, Chapman and Hall/CRC. p. 95-107.

33. Embretson, S.E. and S.P. Reise, *Item response theory*. 2013: Psychology Press.

34. Baker, F.B., *The basics of item response theory*. 2001: ERIC.

35. Teresi, J.A., *Overview of quantitative measurement methods: Equivalence, invariance, and differential item functioning in health applications*. Medical care, 2006. **44**(11): p. S39-S49.

36. Teresi, J.A., *Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics*. Medical care, 2006. **44**(11): p. S152-S170.

37. Hays, R.D., et al., *Differential item functioning by language on the PROMIS(®) physical functioning items for children and adolescents*. Qual Life Res, 2018. **27**(1): p. 235-247.

38. Condon, D.M., et al., *Does recall period matter? Comparing PROMIS(®) physical function with no recall, 24-hr recall, and 7-day recall*. Qual Life Res, 2020. **29**(3): p. 745-753.

39. Choi, S.W., L.E. Gibbons, and P.K. Crane, *Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations*. Journal of statistical software, 2011. **39**(8): p. 1.

40. Edwards, V.J., et al., *Characteristics and health status of informal unpaid caregivers—44 States, District of Columbia, and Puerto Rico, 2015–2017*. Morbidity and Mortality Weekly Report, 2020. **69**(7): p. 183.

41. Streiner, D.L., G.R. Norman, and J. Cairney, *Health measurement scales: a practical guide to their development and use*. 2024: Oxford university press.

42. Acquadro, C., et al., *Literature review of methods to translate health-related quality of life questionnaires for use in multinational clinical trials*. Value in Health, 2008. **11**(3): p. 509-521.

43. Widaman, K.F., E. Ferrer, and R.D. Conger, *Factorial invariance within longitudinal structural equation models: Measuring the same construct across time*. Child development perspectives, 2010. **4**(1): p. 10-18.

Table 1. Participant Characteristics

	Caregivers	**Care recipient/ Patients**
Sociodemographic characteristics (n=504)	**Frequency (%)**	**Frequency (%)**
Age (years), mean (± SD)	49.2 (15.4)	62.7 (18.9)
Age group (years)		
18 – 44	226 (45.2)	102 (20.2)
45 – 64	164 (32.5)	114 (26.6)
65 +	114 (22.6)	288 (57.1)
Gender		
Male	213 (42.3)	238 (47.22)
Female	290 (57.5)	264 (52.38)
Agender (self-described)	1 (0.2)	2 (0.4)
Race/Ethnicity *		
White	369 (73.2)	362 (71.8)
Black or African American	79 (15.7)	79 (15.7)
American Indian or Alaskan Native	13 (2.6)	5 (0.9)
Asian	27 (5.4)	26 (5.1)
Hispanic or Latino or Spanish Origin of any race	62 (12.3)	55 (10.9)
Native Hawaiian/Other Pacific Islander	1 (0.2)	0 (0)
Multi-racial	3 (0.6)	3 (0.6)
Employment status		
Employed (full-time, part-time or self-employed)	311 (61.7)	58 (11.5)
Retired, homemaker	138 (27.4)	233 (46.2)
Student, unemployed (unable to work due to disability, looking or not looking for work)	55 (10.9)	213 (42.3)
Marital status		
Married, engaged, living with partner	350 (69.4)	248 (49.2)
Widowed, divorced or separated	71 (14.1)	184 (31.7)
Single, never married	83 (16.5)	72 (36.9)
Educational attainment		
High school degree/GED or less	103 (20.4)	229 (45.4)
Technical school, associate or some college (no degree)	204 (40.5)	131 (26.0)
Bachelor's degree	106 (21.0)	86 (17.1)
Master's, professional or doctorate degree	91 (18.1)	58 (11.5)

Difficulty in meeting monthly household expenses		
Not difficult	195 (38.7)	206 (40.87)
Slightly difficult	146 (29.0)	125 (24.8)
Somewhat difficult	81 (16.1)	83 (16.47)
Very difficult	52 (10.3)	55 (10.91)
Extremely difficult	30 (6.0)	35 (6.94)
Health and relationship quality	**Frequency (%)**	**Frequency (%)**
Health and well-being measures		
EQ-5D-5L Index	0.73 (0.28)	0.43 (0.40)
EQ VAS	71.45 (20.63)	55.33 (23.66)
EQ-HWB-S Index	0.67 (0.26)	0.47 (0.03)
Caregiver burden		
ASCOT Index	0.72 (0.23)	
Carer-QoL Index	70.28 (21.86)	
Carer-QoL VAS	6.72 (2.26)	
Care recipient perceived burden		
CARE-SOB scale total		13.27 (5.89)
SPB-scale summary score (± SD)		28.50 (9.60)
SPB-scale: little burden, n (%)		106 (21.0)
SPB-scale: mild to moderate burden		163 (32.3)
SPB-scale: moderate to severe burden		160 (31.8)
SPB-scale: very severe burden		75 (14.9)
Caregiving Situation	**Frequency (%)**	
Relationship to care recipient		
Spouse/Partner	174 (34.5)	
Parent	21 (4.2)	
Child	150 (29.8)	
Sibling	31 (6.2)	
Another relative (not child, sibling, parent, grandparent)	30 (6.0)	
Friend/Family Friend	60 (11.9)	
Grandchild	38 (7.5)	
Reason for providing assistance to care recipient		
Physical condition (short term)	75 (14.9)	
Physical condition (long-term)	297 (58.9)	
Emotional or mental health problem	148 (29.4)	

Developmental or intellectual disability or delay	35 (6.9)
Behavioral issue	50 (9.9)
Memory problem	127 (25.2)
Old age, aging	237 (47.0)
Other	31 (6.2)
Duration of caregiving (years)	
6 months – 1	48 (9.5)
1 – 2	141 (28.0)
3 – 5	158 (31.4)
6 – 10	88 (17.5)
10 >	69 (13.7)
Primary caregiver	
Yes	439 (87.1)
No	10 (2.0)
Sharing caregiving responsibilities about equally with someone else	55 (10.9)
Level of Care Index (intensity of caregiving)*	
Level 1	27 (5.4)
Level 2	38 (7.5)
Level 3	87 (17.3)
Level 4	274 (54.4)
Level 5	78 (15.5)
Average weekly time spent on caregiving (hours)	
> 5	35 (6.9)
6 – 10	69 (13.7)
11 – 20	176 (34.9)
21 – 30	99 (19.6)
31 – 40	28 (5.6)
40 >	46 (9.1)
Living in the same household as care recipient (yes)	344 (68.3)

** The Level of Care Index, comprising five levels, categorizes caregivers based on caregiving intensity by combining hours of care per week and types of care provided (IADLs and ADLs). Level 1 signifies the least intense caregiving, while Level 5 represents the most intense caregiving.*

The Burns Relationship Scale is a self-report measure used to assess the quality of interpersonal relationships, focusing here on the caregiver-care recipient dyad. It examines aspects such as communication, trust, and emotional closeness. Higher scores on the scale indicate a stronger, more positive relationship between the caregiver and care recipient.

EQ-5D-5L, and Carer-QoL (The Carer-related Quality of Life) index scores were calculated using US-specific utility values, and EQ-HWB-S used pilot data utility values for the UK. Higher scores on EQ-5D-5L, EQ VAS, EQ-HWB-S, and CarerQoL represent better health and quality of life.

Table 2. Response Distribution of Single-Domain and Double-Barreled Items in Caregivers and Patients

	Caregivers				Patients			
	Concentrating	Thinking Clearly	Concentrating/ Thinking Clearly	Difference	Concentrating	Thinking Clearly	Concentrating/ Thinking Clearly	Difference
	N (%)	**N (%)**	**N (%)**		**N (%)**	**N (%)**	**N (%)**	
None of the time	175 (34.72%)	211 (41.87%)	168 (33.33%)	-4.96%	86 (17.06%)	95 (18.85%)	72 (14.29%)	-3.67%
Only occasionally	176 (34.92%)	165 (32.74%)	158 (31.35%)	-2.48%	137 (27.18%)	155 (30.75%)	132 (26.19%)	-2.78%
Sometimes	83 (16.47%)	83 (16.47%)	102 (20.24%)	+3.77%	154 (30.56%)	150 (29.76%)	143 (28.37%)	-1.79%
Often	51 (10.12%)	34 (6.75%)	61 (12.10%)	+3.67%	86 (17.06%)	74 (14.68%)	115 (22.82%)	+6.95%
Most of the time	19 (3.77%)	11 (2.18%)	15 (2.98%)	0.00%	41 (8.13%)	30 (5.95%)	42 (8.33%)	+1.29%
	Walking Inside	Walking Outside	Walking Inside/Outside	Difference	Walking Inside	Walking Outside	Walking Inside/ Outside	Difference
	N (%)	**N (%)**	**N (%)**		**N (%)**	**N (%)**	**N (%)**	
No Difficulty	317 (62.90%)	286 (56.75%)	303 (60.12%)	+0.30%	107 (21.23%)	85 (16.87%)	93 (18.45%)	-0.60%
Slight Difficulty	106 (21.03%)	109 (21.63%)	98 (19.44%)	-1.89%	118 (23.41%)	97 (19.25%)	114 (22.62%)	+1.29%
Some Difficulty	60 (11.90%)	74 (14.68%)	75 (14.88%)	+1.59%	157 (31.15%)	126 (25.00%)	141 (27.98%)	-0.10%
A lot of difficulty	20 (3.97%)	27 (5.36%)	23 (4.56%)	-0.10%	95 (18.85%)	141 (27.98%)	132 (26.19%)	+2.78%
Unable	1 (0.20%)	8 (1.59%)	5 (0.99%)	+0.10%	27 (5.36%)	55 (10.91%)	24 (4.76%)	-3.37%

Table 3. Correlation Coefficients and Level of Agreement for Caregivers and Patients across Concentrating/Thinking Clearly and Walking Inside/Outside Items

	Concentrating/ Thinking Clearly			Walking Inside/Outside	
Correlation Analysis	Caregivers	Patients		Caregivers	Patients
	r_s	r_s		r_s	r_s
Concentrating	0.78	0.74	Walking Inside	0.70	0.75
Thinking Clearly	0.76	0.74	Walking Outside	0.75	0.71
Level of Agreement	Caregivers	Patients		Caregivers	Patients
	κ	κ		κ	κ
Concentrating	0.79	0.74	Concentrating	0.69	0.76
Thinking Clearly	0.75	0.72	Thinking Clearly	0.72	0.71

Note: all $p<0.05$; κ: Cohen's kappa coefficient; r_s: Spearman's rank correlation coefficient

Table 4. Factor Loadings from Confirmatory Factor Analysis for Caregivers and Patients

		Caregiver	Patient
Factor	**Item**	**Factor Loading**	**Factor Loading**
Cognition	Concentrating	0.951	0.920
	Thinking Clearly	0.954	0.967
	Conc./Thinking Clearly (double)	0.880	0.913
Mobility	Walking Inside	0.974	0.904
	Walking Outside	0.954	0.976
	Walking In/Out (double)	0.877	0.948
Factor Correlation	Cognition with Mobility	0.552	0.375
		Model Fit Indices	**Model Fit Indices**
	CFI	1	0.999
	TLI	1	0.997
	RMSEA (90% CI)	0.004 (0.000-0.037)	0.097 (0.079-0.116)
	SRMR	0.006	0.014

CFI: Comparative Fit Index, TLI: Tucker-Lewis Index, RMSEA (90% CI): Root Mean Square Error of Approximation (90% Confidence Interval), SRMR: Standardized Root Mean Square Residual

Table 5. Item Response Theory Parameter Estimates for Single-Domain and Double-Barreled Items

	Caregiver			Patient		
	Concentrating	Thinking Clearly	Concentrating/ Thinking Clearly	Concentrating	Thinking Clearly	Concentrating/ Thinking Clearly
Threshold 1	-0.38	-0.19	-0.45	-0.98	-0.90	-1.23
Threshold 2	0.53	0.68	0.42	-0.11	0.01	-0.24
Threshold 3	1.12	1.42	1.13	0.69	0.83	0.57
Threshold 4	1.80	2.05	2.07	1.43	1.60	1.58
Slope	6.59	5.49	3.62	5.78	5.90	3.01
	Walking Inside	Walking Outside	Walking Inside/ Outside	Walking Inside	Walking Outside	Walking Inside/ Outside
Threshold 1	0.36	0.18	0.31	-0.78	-1.03	-1.01
Threshold 2	1.04	0.84	0.96	-0.09	-0.32	-0.23
Threshold 3	1.75	1.52	1.82	0.68	0.34	0.54
Threshold 4	2.91	2.12	2.64	1.61	1.28	1.90
Slope	5.68	6.65	3.13	8.63	4.22	3.10

Table 6. Differential Item Functioning (DIF) Analysis for Single-Domain and Double-Barreled Items for Caregivers and Patients

		Caregiver			Patient		
		Uniform DIF	**Non-uniform DIF**	**Overall DIF**	**Uniform DIF**	**Non-uniform DIF**	**Overall DIF**
Item 1	**Item 2**	ΔR^2	ΔR^2	ΔR^2	ΔR^2	ΔR^2	ΔR^2
Thinking Clearly	Conc./ Thinking Clearly	0.0035	0.0011	0.0047	0.0004	0.0052	0.0056
Concentrating	Conc./ Thinking Clearly	0	0.0007	0.0007	0	0.0079	0.0079
Walking Outside	Walking Outside/Inside	0.0003	0	0.0003	0	0	0
Walking Inside	Walking Outside/Inside	0.0012	0.0001	0.0013	0.0241	0.0068	0.0309

Uniform DIF: Uniform Differential Item Functioning, Non-uniform DIF: Non-uniform Differential Item Functioning, Overall DIF: Overall Differential Item Functioning, R^2: McFadden's R-squared

Figure 1. Item Characteristic Curves (ICC) and Item Information Traces (IIT) for Caregivers and Patients: Concentrating/Thinking Clearly

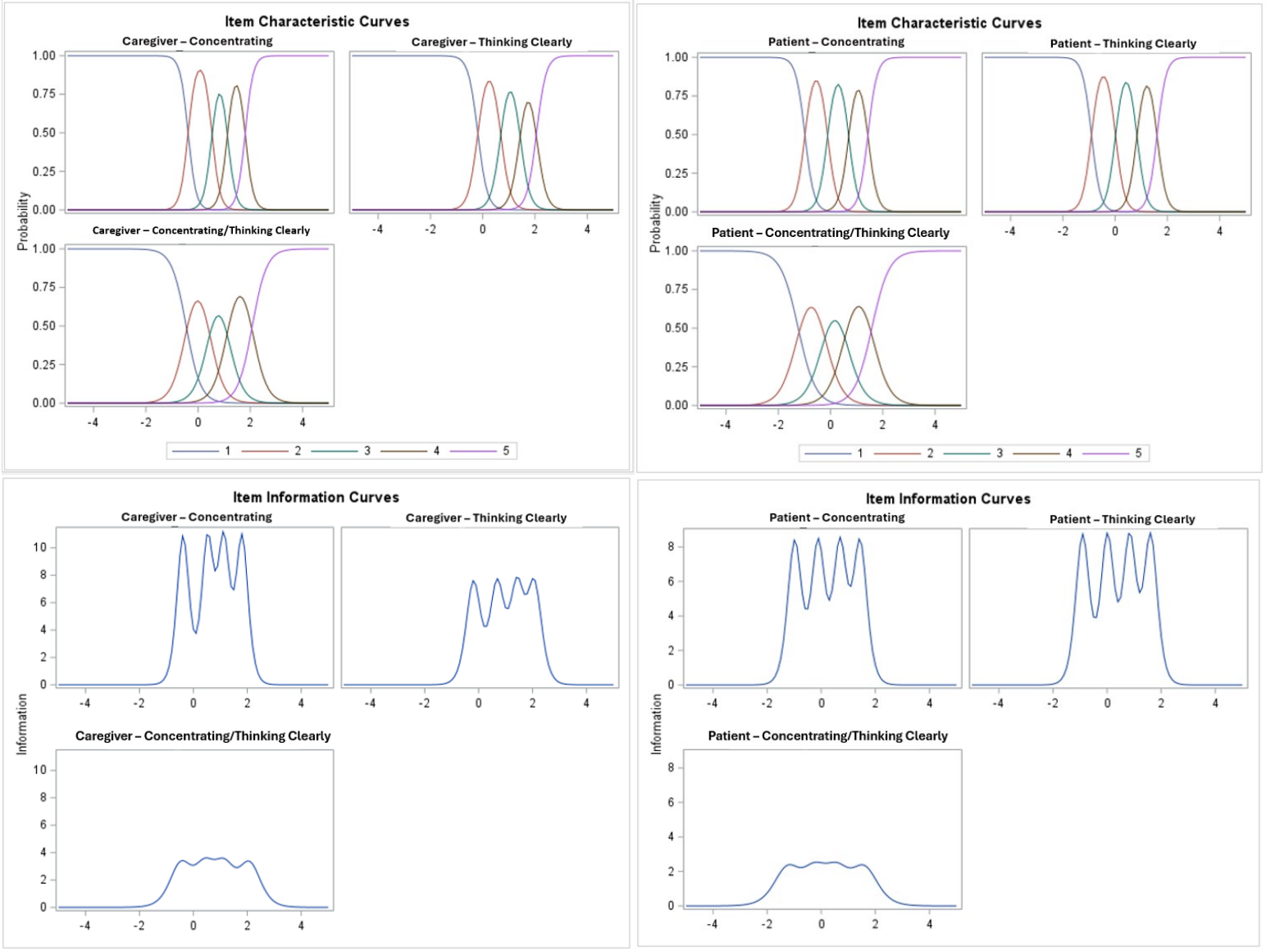

Figure 2. Item Characteristic Curves (ICC) and Item Information Traces (IIT) for Caregiver and Patient: Walking Inside and Outside