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

Patient-Reported Outcomes

An Investigation of Age-Related Differential Item Functioning in the EQ-5D-5L Using Item Response Theory and Logistic Regression

Hannah Penton, PhD, Christopher Dayson, MA, Claire Hulme, PhD, Tracey Young, PhD

ABSTRACT

Objectives: In economic evaluations, quality of life is measured using patient-reported outcome measures (PROMs), such as the EQ-5D-5L. A key assumption for the validity of PROMs data is measurement invariance, which requires that PROM items and response options are interpreted the same across respondents. If measurement invariance is violated, PROMs exhibit differential item functioning (DIF), whereby individuals from different groups with the same underlying health respond differently, potentially biasing scores. One important group of healthcare consumers who have been shown to have different views or priorities over health is older adults. This study investigates age-related DIF in the EQ-5D-5L using item response theory (IRT) and ordinal logistic regression approaches.

Methods: Multiple-group IRT models were used to investigate DIF, by assessing whether older adults aged 65+ years and younger adults aged 18 to 64 years with the same underlying health had different IRT parameter estimates and expected item and EQ-5D-5L level sum scores. Ordinal logistic regression was also used to examine whether DIF resulted in meaningful differences in expected EQ level sum scores. Effect sizes examined whether DIF indicated meaningful score differences.

Results: The anxiety/depression item exhibited meaningful DIF in both approaches, with older adults less likely to report problems. Pain/discomfort and mobility exhibited DIF to a lesser extent.

Conclusions: When using the EQ-5D-5L to evaluate interventions and make resource allocation decisions, scoring bias due to DIF should be controlled for to prevent inefficient service provision, where the most cost-effective services are not provided, which could be detrimental to patients and the efficiency of health budgets.

Keywords: differential item functioning, EQ-5D, item response theory, ordinal logistic regression, patient-reported outcome measures, quality of life, response heterogeneity.

VALUE HEALTH. 2022; ■(■):■–■

Introduction

Quality-adjusted life-years (QALYs) are a commonly used outcome measure in the economic evaluation of health interventions.[1] QALYs combine length of life and a utility value for quality of life (QOL) into a single unit, enabling comparisons of different interventions across different patient groups for health service resource allocation. Given that QOL cannot be directly observed, preference-based patient-reported outcome measures (PROMs) are used to measure and value QOL. The EQ-5D measure of health is the most widely used PROM and is recommended or required to measure utility by multiple Health Technology Assessment agencies including the UK National Institute for Health and Care Excellence, the Dutch Zorginstituut, and the Canadian Agency for Drugs and Technologies in Health.[2-4]

A key assumption, upon which the validity of any PROM data depends, is measurement invariance.[5] Measurement invariance requires that the concept being measured by a PROM and the questions and response options within a PROM mean the same thing across different groups of respondents. If this assumption does not hold, the PROM exhibits differential item functioning (DIF) (also called response heterogeneity), whereby individuals from different subgroups with the same underlying levels of health have different probabilities of providing a given response to an item on that PROM, which will bias their score. There are 2 types of DIF: uniform and nonuniform. Uniform DIF occurs when groups have different probabilities of providing a given response at a given level of trait, and nonuniform DIF occurs when the extent to which the items relate to the underlying trait differs.[6] If DIF is present, comparisons between different respondents are invalid, as would be any decisions based on such comparisons. Given the wide use of such PROMs in population health studies and economic evaluation, DIF could lead to bias in estimates of burden of illness, treatment effectiveness, and resource allocation decision making. This may result in inefficient service provision, where the most cost-effective services are not provided, which could be detrimental to patients and the efficiency of health budgets.

An important characteristic in the testing of measurement invariance is age. Older adults, often defined as those ≥ 65 years,[7,8] often experience increasing frailty as they age, characterized by a gradual decline in health and functioning.[9] This gradual decline can lead people's interpretation and prioritization of concepts related to health to change substantially over time, as has been observed in the literature.[9,10] This decline in health and functioning also makes older adults an intensive group of healthcare users. It is estimated that 58% of those aged 65 to 74 years and 68% of those aged 75+ years in Great Britain live with a long-standing illness,[11] and of 18.7 million adult hospital admissions in England in 2014 to 2015, 41% were aged 65+ years.[12] With the population aging and increases in life expectancy outperforming increases in healthy life expectancy, this issue will only intensify with time.[13,14]

With older adults representing such an important group in healthcare spending, correctly estimating the most cost-effective services for this group is a priority for the economic efficiency and sustainability of services in the future. Nevertheless, older adults are often overlooked in measure development and psychometric testing.[9,10] Therefore, this study aims to investigate age-related DIF in the EQ-5D-5L using 2 methods: item response theory (IRT) and ordinal logistic regression (OLR).

Methods

EQ-5D-5L

The EQ-5D-5L measure of health asks individuals to self-report their health today on 5 items: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression.[15] Each question has 5 response options including no problems, slight problems, moderate problems, severe problems, and extreme problems/unable to. The EQ-5D-5L can be scored in several different ways. Most commonly, responses are combined with a country-specific value set to generate utility values for the QALY. Nevertheless, for methodological research examining response behavior, the EuroQol recommend using the simple level sum score (LSS), to isolate response behavior from preferences.[16] This approach will be used in this study to examine DIF without results being affected by preferences for the different items and levels.

To ease interpretation, all items were coded so that higher numbered responses indicated better health. This involved reverse coding all EQ-5D items. Therefore, higher LSSs imply better health in these analyses, contrary to the usual LSS for the EQ-5D.

Data Sources

The Health Improvement and Patient Outcomes dataset is a large patient dataset that collected health and wellbeing data, including the EQ-5D-5L, from inpatients recently discharged from Cardiff and Vale NHS Hospital Trust in 2013 to 2014 via postal survey sent to 25 919 patients aged 18+ years, 6 weeks after discharge.[17] A 25% response rate was obtained with 6351 questionnaires returned.

Analyses

A variety of methods are available for identifying DIF, including IRT, logistic regression, and Mantel-Haenszel (MH) techniques, each with their own advantages and disadvantages.[6,18] The main advantage of IRT is the use of the latent trait as the matching criterion rather than the observed sum score used to proxy the latent trait in logistic regression and MH. The disadvantages of IRT are that results are dependent on model fit, large sample sizes are required, and the procedure is more complex than other methods. MH is nonparametric, so it does not require a distribution assumption and is relatively straightforward, but it is unable to detect nonuniform DIF. Parametric approaches have also been found to be more powerful and stable than nonparametric.[18] Logistic regression is also simpler than IRT and is able to detect both uniform and nonuniform DIF. DIF was tested using 2 methods (IRT and OLR) given that both enable the detection of both uniform and nonuniform DIF and the use of 2 methods increases confidence in results, which are neither entirely dependent on IRT model fit nor the use of the observed sum score to proxy the latent trait.

IRT Analyses

IRT is a class of statistical techniques commonly used in psychometrics to develop measures and assess measure and item performance.[6] It uses responses to PROM items to estimate unobservable health on a latent scale. Logistic models are used to estimate parameters representing the location of respondents and items on this latent scale by examining the probability of a specific item response as a function of the respondent's level of latent health and item characteristics.[19]

IRT models assess item performance using difficulty and discrimination parameters.[20] Discrimination parameters examine how closely an item is related to the underlying health of respondents. An item with n response options also has n-1 difficulty parameters. Each difficulty parameter represents the amount of underlying health required to have a 50% probability of responding above a certain category, signifying better health. For example, given an item with response options 1 to 4, a $b1 = -2$ indicates that someone 2 SDs below mean health has a 50% probability of responding in category 1 and a 50% probability of responding in categories 2 to 4. Therefore, difficulty parameters assess over what levels of health the item can discriminate. Response distributions were also examined for floor and ceiling effects to further signal issues with discrimination.

Polytomous ordinal IRT models were used because all items have > 2 ordered response options. Samejima's 2-parameter graded response model (GRM) was selected, because it has been extensively used in the psychometric evaluation of health measures[21] and allows discrimination parameter estimates to vary across items, which makes theoretical sense given that items within a measure are not equally strongly related to respondents' health.

IRT models can be unidimensional, assuming all items relate to a single latent concept, or multidimensional with items representing one of several constructs. Dimensionality was examined using exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). EFA was used to establish the potential number of factors underlying the measure. The Kaiser rule suggests retaining factors with eigenvalues ≥ 1,[22] and the screen test suggests that the number of true factors ends where a line drawn through all the points becomes linear and flat.[23] CFA was run, and absolute model fit compared between potentially appropriate models based on EFA results. Model fit was examined using the root mean square error of approximation (RMSEA) and comparative fit index (CFI). Cutoffs for good model fit are CFI $\geq$ 0.95 and RMSEA $\leq$ 0.05, with RMSEA $\leq$ 0.08 considered fair.[24]

IRT models assume local independence, meaning there is no additional systematic covariance between items beyond their given relationship with the underlying trait.[25] Local dependence (LD) may arise in groups of items with similar content or that are physically grouped together on a measure. Groups of items with large modification indices of error covariances or excessively high discrimination parameter estimates or interitem correlations may exhibit LD.[25]

IRT analyses were run in Mplus version 7.4.[26] DIF was investigated using multiple-group GRMs that allow separate parameter estimates for respondents aged 18 to 64 years (reference group) and 65+ years (focal group) to investigate how psychometric performance varies between these groups.[27] Models were estimated using partial information weighted least squares with mean and variance adjustment (WLSMV) with theta parameterization to obtain absolute fit statistics. WLSMV was chosen because it does not assume the underlying data follow a normal distribution and provides absolute fit statistics and easy comparison of nested model fit through the Mplus DIFFTEST function. WLSMV estimation in Mplus provides parameter estimates consistent with categorical normal-ogive CFA, but IRT parameters consistent with GRMs were estimated through transformations.

A step-by-step process for DIF testing within IRT has been developed and widely used.[5,27-29] DIF is tested one by one in factor structure (configural invariance); discrimination parameter estimates, to examine whether items are equally related to latent health across age groups (violation signals nonuniform DIF); difficulty parameter estimates, to investigate whether different groups have the same probability of selecting each response option at given levels of underlying health (violation signals uniform DIF); and residual variances, factor means, and factor variances. The model specification at each stage of DIF testing is outlined in Appendix Table A1 in Supplemental Materials found at https://dx.doi.org/10.1016/j.jval.2022.03.009. Relative nested model fit to the final model in the previous DIF stage was compared using the DIFFTEST function in Mplus.

Although this method indicates where significant DIF lies, it does not quantify the impact of DIF on scores.[30] The magnitude of DIF can be examined using expected item scores for each group, calculated by summing the weighted probability of each response (weighted by the response category value) at each level of underlying health from the final DIF model. Differences in expected item and total measure scores provide an estimate of the impact of DIF at the item and measure level, at each level of health. The impact of DIF is presented as a percent of the total LSS range. The authors again note that the LSS does not reflect the standard final scoring method for the EQ-5D-5L. Differences in LSS are reflective of the raw expected difference in reporting rather than the final expected difference in utility. Effect sizes were also calculated to aid interpretation of whether differences in expected scores between groups were of practical importance. Expected score standardized differences (ESSD) for each item were estimated after the procedure outlined by Meade.[31] ESSDs were classified as trivial (ESSD < 0.2), small ($0.2 \leq$ ESSD < 0.5), moderate ($0.5 \leq$ ESSD < 0.8), or large (ESSD ≥ 0.8).[31]

The full sample was randomly split in STATA into a model development and validation sample. Differences between samples were investigated using the 2-sample Wilcoxon rank-sum (Mann-Whitney) test for continuous variables and Pearson's chi-square test for categorical variables. The final DIF model obtained from the development sample was rerun in the validation sample. Item parameter estimates, expected scores, and DIF impact were compared across samples to examine the stability of results found in the development sample.

OLR Analyses

The OLR approach for detecting DIF proposed by Zumbo[32] was followed. In the primary OLR analysis, the purified EQ LSS was used as the matching variable, to proxy the latent trait, because the EuroQol recommend the use of the LSS when analyzing descriptive system issues, to isolate response behavior from preferences.[16] The LSS was purified by removing one by one any items that exhibited meaningful DIF from the LSS. For the final analysis of an item which exhibited DIF, the item in question was added back into the LSS. There has been concern in the literature that purification may not be appropriate in shorter scales because it leaves too few items on which to match and proxy the latent trait.[18] Therefore, a sensitivity analysis was also conducted in which the LSS was not purified.

Logistic regression models were estimated, with item score as the dependent variable and EQ LSS, age group, and an interaction for LSS and age group as dependent variables, as shown in Eq. (1).

$$Y = b_o + b_1 LSS + b_2 AGE\_GROUP + b_3 LSS * AGE\_GROUP \quad (1)$$

The OLR DIF approach involves an iterative 3-stage process enabling the detection of both uniform and nonuniform DIF. In stage 1, an OLR is run with variable item score as the dependent variable and EQ LSS as the only independent variable (model 1). In stage 2, the age group (18-64 = 0, 65+ = 1) independent variable is added to model 1. If the age group coefficient is statistically significant (assuming a Bonferroni corrected P value of .05/5 items = .01), then this item is considered to exhibit uniform DIF. In stage 3, the observed health × age group interaction is added to model 2 (model 3). If the interaction is statistically significant, then the item is considered to exhibit nonuniform DIF.

As in the IRT approach, a measure of effect size is also required to avoid overinterpretation of practically meaningless DIF. A variety of effect size measures are used in the literature.[18] This study used the approach suggested by Bjorner et al,[33] in which model 3 is compared with model 1 using Nagelkerke's pseudo-R^2 statistic. Items with significant P values, for which $\Delta R^2 \geq 2\%$, are considered to exhibit meaningful DIF. This approach has been widely used in the literature.[33-35] Differences in R^2 between models 1 and 2 and models 2 and 3 can also be used to further examine whether the majority of the DIF stems from uniform or nonuniform DIF.

OLR analyses were conducted in SPSS version 27.[36] As recommended, observations with missing EQ-5D data were removed before analysis.

Results

Sample

Sample characteristics are presented in Appendix Table A2 in Supplemental Materials found at https://dx.doi.org/10.1016/j.jval.2022.03.009. The IRT model development sample contains 1818 adults aged 18 to 64 years and 1333 adults aged 65+ years who provided responses to at least some EQ-5D items. Differences in characteristics between the development and validation sample were not significant in all characteristics tested.

Response Distributions

Response distributions are presented in Table 1. Younger adults reported few problems with self-care, anxiety/depression, mobility, and usual activities, and older adults reported few problems with self-care and anxiety/depression. Younger adults were more likely to report no problems to mobility, self-care,

Table 1. EQ-5D-5L response distributions.

Development sample												
Younger than 65 years old, n = 1818, n (%)							65+ years old, n = 1333, n (%)					
1 Extreme	2 Severe	3 Moderate	4 Slight	5 None	Missing	Item	1 Extreme	2 Severe	3 Moderate	4 Slight	5 None	Missing
19 (1.0)	173 (9.5)	260 (14.3)	277 (15.2)	1075 (59.1)	14 (0.8)	Mobility	22 (1.7)	218 (16.4)	340 (25.5)	254 (19.1)	490 (36.8)	9 (0.7)
7 (0.4)	58 (3.2)	153 (8.4)	220 (12.1)	1374 (75.6)	4 (0.3)	Self-care	24 (1.8)	56 (4.2)	167 (12.5)	181 (13.6)	892 (66.9)	13 (1.0)
85 (4.7)	164 (9.0)	299 (16.4)	392 (21.6)	869 (47.8)	9 (0.5)	Usual activities	74 (5.6)	164 (12.3)	317 (23.8)	335 (25.1)	435 (32.6)	8 (0.6)
62 (3.4)	187 (10.3)	410 (22.6)	607 (33.4)	541 (29.8)	11 (0.6)	Pain/ discomfort	23 (1.7)	175 (13.1)	372 (27.9)	432 (32.4)	319 (23.9)	12 (0.9)
33 (1.8)	91 (5.0)	249 (13.7)	456 (25.1)	979 (53.9)	10 (0.6)	Anxiety/ depression	9 (0.7)	27 (2.0)	188 (14.1)	327 (24.5)	767 (57.5)	15 (1.1)

Validation sample												
Younger than 65 years old, n = 1793, n (%)							65+ years old, n = 1356, n (%)					
1 Extreme	2 Severe	3 Moderate	4 Slight	5 None	Missing	Item	1 Extreme	2 Severe	3 Moderate	4 Slight	5 None	Missing
16 (0.9)	164 (9.1)	230 (12.8)	299 (16.7)	1073 (59.8)	11 (0.6)	Mobility	23 (1.7)	205 (15.1)	351 (25.9)	246 (18.1)	522 (38.5)	9 (0.7)
19 (1.1)	43 (2.4)	160 (8.9)	172 (9.6)	1389 (77.5)	10 (0.6)	Self-care	20 (1.5)	44 (3.2)	168 (12.4)	227 (16.7)	886 (65.3)	11 (0.8)
72 (4.0)	153 (8.5)	290 (16.2)	410 (22.9)	861 (48.0)	7 (0.4)	Usual activities	77 (5.7)	142 (10.5)	333 (24.6)	332 (24.5)	465 (34.3)	7 (0.5)
71 (4.0)	189 (10.5)	369 (20.6)	584 (32.6)	568 (31.7)	12 (0.7)	Pain/ discomfort	25 (1.8)	173 (12.8)	390 (28.8)	433 (31.9)	321 (23.7)	14 (1.0)
50 (2.8)	88 (4.9)	266 (14.8)	453 (25.3)	928 (51.8)	8 (0.4)	Anxiety/ depression	8 (0.6)	29 (2.1)	175 (12.9)	369 (27.2)	754 (55.6)	21 (1.5)

Note. All items are coded so that higher numbered categories signal better health.
EQ-5D-5L indicates 5-level version of EQ-5D.

usual activities, and pain than older adults but less likely to report no problems to anxiety/depression.

IRT Assumption Checks

The EFA eigenvalues (3.8 and 0.58) and scree plot suggested a 1-factor model. The 2-factor EFA loadings indicated that only anxiety/depression would load onto the second factor. Therefore, a single factor model was taken forward. Modification indices did not suggest any LD.

IRT Results

Absolute fit statistics for the multiple-group GRM were good (CFI 0.999; RMSEA 0.047 [90% confidence interval 0.037-0.058]). Unstandardized discrimination parameter estimates (Table 2) ranged from 2.64 for mobility to 0.86 for anxiety/depression for both age groups, suggesting mobility is most closely related to health in both age groups and anxiety/depression least related. Nevertheless, anxiety/depression is still relevant to health with the smallest standardized discrimination parameter (0.634) in respondents aged 65+ years. Larger discriminations for pain/discomfort and self-care in those younger than 65 years indicate that these concepts are less closely related to the health of respondents aged 65+ years than those younger than 65 years. This could be anticipated because those aged 65+ years may be more accustomed to issues in self-care and pain and may have adapted.

The difficulty parameter estimates (b1-b4; Table 2) represent the amount of health required to have a 50% probability of responding above a certain category, signaling better health. All difficulty parameter estimates for pain/discomfort and anxiety/depression were lower in respondents aged 65+ years, meaning those aged 65+ years require less health to be more likely to respond in a higher category (signaling better health) than younger adults. Therefore, older adults are less likely to report problems with pain/discomfort and anxiety/depression than a younger person with the same level of health. Older adults were also more likely to respond higher (signaling less problems) to self-care and usual activities but lower to mobility than a younger adult with the same health level.

Anyone with even slightly below average health is more likely to respond "no problems" for self-care, anxiety/depression, and mobility because these items all have negative b4 parameter estimates. These low b4s correspond to the high endorsement rate of "no problems" for these items. Anxiety/depression also has very low b1 parameter estimates, below −3, in both groups, with the b2 parameter also below −3 in the 65+ group. Theoretically, we would expect very few respondents to have a health level below 3 SDs below mean health, suggesting the "extreme" and "severe" anxiety/depression categories would very rarely be used, particularly by older respondents. This was found to be true in the empirical item response frequency distributions.

Table 2. IRT parameter estimates.

Younger than 65 years old						65+ years old				
a	b1	b2	b3	b4	Item	a	b1	b2	b3	b4
2.64 (0.12)	−2.51 (0.21)	−1.31 (0.15)	−0.72 (0.12)*	−0.26 (0.09)*	Mobility	2.64 (0.12)	−2.51 (0.21)	−1.31 (0.15)	−0.52 (0.11)*	−0.02 (0.11)*
2.47 (0.13)*	−2.87 (0.35)*	−1.94 (0.21)*	−1.27 (0.16)*	−0.76 (0.13)*	Self-care	1.86 (0.10)*	−2.66 (0.17)*	−2.06 (0.16)*	−1.33 (0.13)*	−0.86 (0.11)*
2.32 (0.09)	−1.82 (0.13)*	−1.19 (0.11)*	−0.57 (0.09)	0.08 (0.07)	Usual activities	2.32 (0.09)	−2.03 (0.16)*	−1.32 (0.13)*	−0.57 (0.09)	0.08 (0.07)
1.69 (0.05)*	−2.18 (0.10)*	−1.26 (0.06)*	−0.42 (0.05)*	0.61 (0.06)*	Pain/ discomfort	1.30 (0.06)*	−2.83 (0.10)*	−1.64 (0.06)*	−0.55 (0.05)*	0.50 (0.07)*
0.86 (0.03)	−3.21 (0.09)*	−2.28 (0.06)*	−1.26 (0.05)*	−0.16 (0.04)*	Anxiety/ depression	0.86 (0.03)	−4.08 (0.16)*	−3.26 (0.10)*	−1.80 (0.06)*	−0.67 (0.05)*
0					Factor mean	−0.359 (0.04)				
1					Factor variance	0.914 (0.07)				
0.047 (0.037-0.058)					RMSEA (90% CI)	0.047 (0.037-0.058)				
0.999					CFI	0.999				

CFI indicates comparative fit index; CI, confidence interval; DIF, differential item functioning; IRT, item response theory; RMSEA, root mean square error of approximation.
*Parameters that exhibit DIF; a = discrimination parameter, b = difficulty parameters.

Constraining the unstandardized residual variances of both groups to 1 did not significantly impact model fit (P = .72) indicating that the amount of item variance not accounted for by the factor was equivalent across groups. The mean level of health was 0.359 SDs lower (ie, worse) in the 65+ group (P < .000). Those aged 65+ years were less variable in health (factor variance = 0.914 vs 1) than those younger than 65 years, but this was not significant (P = .248).

Older adults are more likely to score higher (ie, better), given the same underlying health, on all items except mobility, where they are more likely to score lower. DIF had a trivial impact on self-care, usual activities, and mobility scores and a small impact on pain/discomfort and anxiety/depression, with the impact on anxiety/depression approaching moderate, according to effect sizes (Table 3). The impact of DIF on the EQ-5D-5L as a whole was small. The maximum difference in expected LSS was 6.73% of the score range, 2 SDs below mean health, with older adults more likely to score higher (ie, better) (Fig. 1).

IRT Validation

Validation results are presented in the Supplemental Appendix found at https://dx.doi.org/10.1016/j.jval.2022.03.009. Expected LSSs across age groups were very similar in the development and validation samples (Appendix Fig. A1 in Supplemental Materials found at https://dx.doi.org/10.1016/j.jval.2022.03.009). DIF effect sizes (Table 3) were mostly consistent, with several classification differences where effect sizes were close to the borders of different classifications.

OLR Results

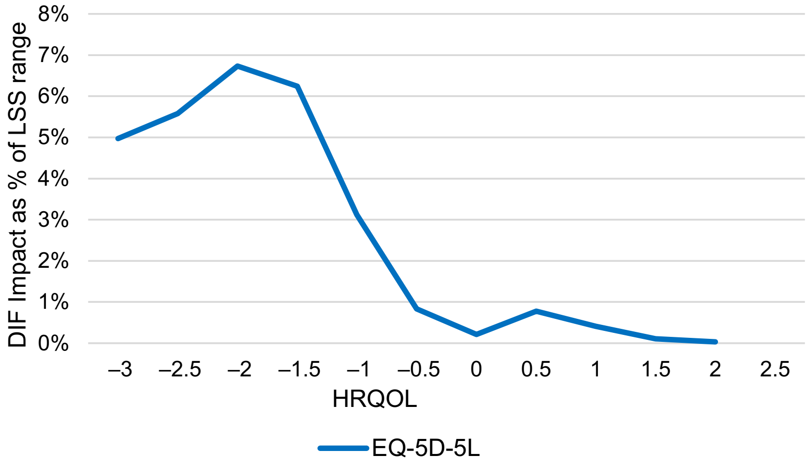
After removing those with missing EQ-5D item data, 3071 participants (1777 aged 18-64 years; 1294 aged 65+ years) remained for analysis in the development sample and 3072 (1758 aged 18-64 years; 1314 aged 65+ years) remained in the validation sample. A total of 4 items had significant age group or interaction variable coefficients indicating DIF: mobility, self-care (development sample only), pain/discomfort, and anxiety/depression (Table 4). The age group coefficients show that older adults were less likely to score highly (indicating better health) on mobility but more likely to signal better health on pain/discomfort and anxiety/depression than adults aged 18 to 64 years. The only item to meet the 2% ΔR^2 cutoff for meaningful DIF was anxiety/depression. The majority of this was uniform DIF. Despite significant coefficients, the impact of that DIF was not large enough to be classed as meaningful for mobility and pain/discomfort.

Table 3. IRT DIF effect sizes in the development and validation samples.

Item	Mobility	Self-care	Usual activities	Pain/discomfort	Anxiety/depression	Total
ESSD development sample	−0.168	0.015	0.059	0.203*	0.475*	0.209*
ESSD validation sample	−0.177	0.022	0.047	0.167	0.550*	0.215*

DIF indicates differential item functioning; ESSD, expected score standardized differences; IRT, item response theory.
*Effect sizes that are small, moderate, or large.

Figure 1. DIF impact as a percent of the LSS range.



DIF indicates differential item functioning; HRQOL, health-related quality of life; LSS, level sum score.

OLR Validation

Validation results (Table 4) mirrored those in the development sample. Results from the unpurified OLR sensitivity analysis (Table 5) show that the impact of DIF on the ΔR^2 for the mobility item was larger in the unpurified analysis, with a ΔR^2 of 0.017 in the development sample and 0.020 in the validation sample, versus 0.009 in the purified analysis. Nevertheless, pain exhibited less DIF in the unpurified analysis compared with the purified, although it did not approach the 2% cutoff in either analysis.

Discussion

Key Findings and Implications

The 2 analysis approaches provide similar results, which increases confidence in results. Meaningful DIF was identified for the anxiety/depression item in both methods and both samples, with older adults less likely to report problems with anxiety/depression than adults aged 18 to 64 years. Pain/discomfort just passed the cutoff for small DIF in the IRT analysis in the development sample; nevertheless, it was not flagged as meaningful in the IRT validation sample or the OLR results (although the ΔR^2 = 0.01 in the validation sample could be considered borderline according to Bjorner[33]), suggesting that overall pain/discomfort could be considered borderline. Mobility exhibited a similar level of DIF to pain/discomfort but was not classified as meaningful (or borderline) in any of the analyses. OLR results and IRT effect sizes identified no/minimal DIF in self-care and usual activities.

Similar results have been seen in the EQ-5D literature. A study in cancer patients using a Rasch partial credit model found large age-related DIF for anxiety/depression and mobility and medium DIF for pain/discomfort and self-care.[37] Another study in individuals undergoing surgical procedures in the UK found age-related DIF anxiety/depression and mobility.[38] Beyond the EQ-5D, DIF studies in other health measures have found that older respondents were less likely to report problems with pain[39] and mental health[30,39-41] and more likely to report problems with

Table 4. OLR results in the development and validation samples.

Item	Model 1	Model 2 (uniform)		Model 3 (nonuniform)		ΔR^2
	R^2	Age coefficient	R^2	Age × LSS coefficient	R^2	
Development sample						
Mobility	0.845	−1.056*	0.854	−0.015	0.854	0.009
Self-care	0.742	−0.128	0.742	0.115*	0.743	0.001
Usual activities	0.836	0.031	0.836	−0.048	0.836	0.000
Pain/discomfort	0.737	0.498*	0.741	−0.114*	0.744	0.007
Anxiety/depression	0.458†	0.799*,†	0.478†	−0.043†	0.480†	0.022†
Validation sample						
Mobility	0.837	−1.026*	0.846	0.015	0.846	0.009
Self-care	0.748	−0.137	0.749	−0.019	0.749	0.001
Usual activities	0.837	0.000	0.837	0.032	0.837	0.000
Pain/discomfort	0.732	0.503*	0.736	−0.156*	0.742	0.010
Anxiety/depression	0.458†	0.918*,†	0.485†	−0.079*,†	0.489†	0.031†

DIF indicates differential item functioning; LSS, level sum score; OLR, ordinal logistic regression.
*$P < .01$.
†Results signal items that met the cutoff for meaningful DIF.

Table 5. OLR results from the unpurified analysis.

Item	Model 1	Model 2 (uniform)		Model 3 (nonuniform)		ΔR²
	R²	Age coefficient	R²	Age × LSS coefficient	R²	
Development sample						
Mobility	0.792	−1.237*	0.809	0.026	0.809	0.017
Self-care	0.714	−0.416*	0.716	0.104*	0.718	0.004
Usual activities	0.805	−0.282*	0.806	−0.013	0.806	0.001
Pain/discomfort	0.727	0.278*	0.729	−0.07*	0.73	0.003
Anxiety/depression	0.458†	0.799*,†	0.478†	−0.043†	0.48†	0.022†
Validation sample						
Mobility	0.777†	−1.266*,†	0.796†	0.063*,†	0.797†	0.020†
Self-care	0.722	−0.567*	0.726	0.029	0.726	0.004
Usual activities	0.805	−0.384*	0.807	0.074*	0.808	0.003
Pain/discomfort	0.722	0.230*	0.723	−0.098*	0.726	0.004
Anxiety/depression	0.458†	0.918*,†	0.485†	−0.079*,†	0.489†	0.031†

DIF indicates differential item functioning; LSS, level sum score; OLR, ordinal logistic regression.
*$P < .01$.
†Results signal items that met the cutoff for meaningful DIF.

physical functioning items such as vigorous or moderate activities[30,39,40] after controlling for underlying health. Similarities in results may signal broader patterns across measures.

These findings align with results from a recent think aloud content validation study in older adults that included the EQ-5D-5L, which could provide an explanation for the results and patterns observed.[42] The content validation study found that older adults were reluctant to signal problems to negatively phrased mental health questions because of a generational stoic attitude of not dwelling on problems. Response shift was also widely observed, with older adults assessing their health relative to people they knew in a worse state or relative to lower expectations of their health in old age. Using these response mechanisms, they lowered their benchmark of what they considered good health and therefore responded more positively than a younger adult would likely respond in the same state. Response shift has also been observed in the wider PROM literature.[43-46] The fact that the statistical DIF findings align with broader literature provides further support to the findings.

These findings present an issue when using the EQ-5D to evaluate interventions and make resource allocation decisions. Bias in scores of different age groups could affect decision making in different ways. Within an evaluation for an intervention aimed at a broad age range of patients, it could cause different age groups to receive inappropriately different estimates of effectiveness. Effectiveness is often measured using self-reported health before and after intervention. It may be expected that if each self-report is consistently higher or lower for a certain group, the effectiveness estimate, the difference between the 2 assessments may not be biased. Nevertheless, this will not necessarily be the case. If older adults are less likely than younger adults to report problems in their preintervention self-report, there is less room on the descriptive system to report improvements, which may result in a smaller estimate of effectiveness than would be observed in younger adults and biased results. If subgroup analysis is conducted, this could result in the intervention only being provided to some individuals whereas others are denied the intervention (which should have been cost-effective), based solely on their age. Conversely, an intervention could be inappropriately approved in a subgroup in which it is not truly cost-effective, leading to a waste of resources. If the intervention is aimed at a single age group, the effectiveness estimates could simply be lower or higher than they should in fact be, potentially leading to similar errors in decision making. At the National Health Service level, interventions that may only be appropriate for certain age groups compete for funding. Therefore, bias in effectiveness estimates could unfairly bias funding decisions for or against certain age groups.

Limitations

The accuracy of DIF detection using IRT is dependent on sample size, model fit, type of IRT model used, choice of anchor items, and dimensionality.[47,48] Sample sizes of ≥ 500 per group are recommended for stable parameter estimates.[47] The smallest group size in this analysis was 1333. A variety of different IRT models were tested, and the best fitting model, which exhibited good fit statistics, was chosen. The standard errors of item parameter estimates can provide an indication of estimation accuracy, with a cutoff of standard error < 0.35 indicating a good level of accuracy.[47] This cutoff was achieved by all items. The similarity between results in the development and validation samples also provides further confidence in the results. Finally, IRT results are somewhat dependent on the choice of anchor items within the analysis. It should be noted that in cases where most items show DIF, the choice of anchor items is somewhat arbitrary, and this should be considered when reflecting on results. Given these limitations, OLR methods were also examined to either confirm the IRT results or examine the extent that misfit and choice of anchor items may have affected results.

Another issue with IRT methods is their high power to detect very small DIF in large samples.[31] Large required samples increase the likelihood of identifying statistically significant, but practically unimportant DIF. Effect size measures were estimated to assist in the interpretation of DIF impact, as recommended in the literature to reduce overinterpretation of DIF that has minimal impact on scores.[31,48]

Another limitation of the analysis relates to the dimensionality assumed and the impact on scoring. The EQ-5D-5L is theorized to be a 5-dimensional measure, with each domain represented by a single item. There is ongoing debate within the literature as to whether the EQ-5D can be treated as unidimensional or whether the 5-dimensional structure must be followed, which would make IRT analyses on the EQ-5D alone infeasible with a single item per domain. Previous factor analyses and IRT analyses have identified unidimensional structures with acceptable fit statistics[49,50]; nevertheless, other studies have found multifactor solutions when the EQ-5D is analyzed in factor analysis alongside other measures.[51-53] Tests of multidimensionality have limited power in a measure of 5 items. Therefore, some of the DIF identified using IRT methods could be due to unaccounted for multidimensionality. This should be considered when reflecting on the IRT results. OLR DIF detection methods, which do not rely on unidimensionality, provide a benchmark of the extent to which multidimensionality has affected the IRT results.

Limitations related to the OLR DIF approach include a lack of consensus on the cutoff for meaningful DIF and the use of the observed LSS to proxy the latent trait. There are a wide variety of different approaches to cut off and classify DIF in logistic regression, with no consensus on the best approach.[18] Different cutoffs lead to very different findings regarding what constitutes meaningful DIF. Another criticism of logistic regression DIF methods is that the observed sum score may not be a good proxy for the latent trait, particularly in short measures that exhibit DIF, where the purification process means that removing items from the LSS to avoid them biasing the results of other items leads to a matching variable based on very few items that provide an even more limited representation of the latent trait. This is why a sensitivity analysis using an unpurified LSS was also conducted.

Again, it is noted that differences in expected scores on the EQ-5D-5L represent the impact of DIF on the LSS, not utilities, given that the use of value set scoring would mix preferences for different items and levels with response behavior that would confuse results.

Conclusions

DIF was identified in the EQ-5D-5L LSS using both IRT and OLR methods of DIF detection. When using this measure to evaluate interventions and allocate resources, bias due to DIF may result in inefficient service provision, where the most cost-effective services are not provided, which could be detrimental to both patients and budget efficiency. Methods to reduce or account for DIF should be further explored.

Supplemental Materials

Supplementary data associated with this article can be found in the online version at https://dx.doi.org/10.1016/j.jval.2022.03.009.

Article and Author Information

Accepted for Publication: March 19, 2022

Published Online: xxxx

doi: https://doi.org/10.1016/j.jval.2022.03.009

Author Affiliations: School of Health and Related Research, University of Sheffield, Sheffield, England, UK (Penton, Young); Centre for Regional Economic and Social Research, Sheffield Hallam University, Sheffield, England, UK (Dayson); Institute of Health Research, University of Exeter, St Luke's Campus, Exeter, England, UK (Hulme).

Correspondence: Hannah Penton, PhD, Open Health, Marten Meesweg 107, 3068 AV, Rotterdam, The Netherlands. Email: hannahpenton@openhealthgroup.com

Author Contributions: *Concept and design:* Penton, Dayson, Hulme, Young
Analysis and interpretation of data: Penton
Drafting of the manuscript: Penton, Dayson, Hulme, Young
Critical revision of the paper for important intellectual content: Penton, Dayson, Hulme, Young
Statistical analysis: Penton
Supervision: Penton, Dayson, Hulme, Young

Conflict of Interest Disclosures: The authors reported no conflicts of interest. The views expressed in this publication are those of the authors and not necessarily those of the National Institute for Health Research or the Department of Health and Social Care.

Funding/Support: This study was funded by the National Institute for Health Research Yorkshire and Humber Applied Research Collaboration.

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

Acknowledgment: This report is independent research funded by the National Institute for Health Research Yorkshire and Humber Applied Research Collaboration.

REFERENCES

1. Drummond M, Sculpher M, Torrance G, O'Brien B, Stoddart G. *Methods for the Economic Evaluation of Health Care Programmes*. 3rd ed. Oxford, United Kingdom: Oxford University Press; 2005.
2. Guide to the methods of technology appraisal 2013. National Institute for Health and Care Excellence. http://www.nice.org.uk/media/D45/1E/GuideToMethodsTechnologyAppraisal2013.pdf. Accessed May 3, 2020.
3. Guideline for Economic Evaluations in Healthcare. Zorginstituut Nederlands. https://english.zorginstituutnederland.nl/publications/reports/2016/06/16/guideline-for-economic-evaluations-in-healthcare. Accessed May 3, 2020.
4. Guidelines for the economic evaluation of health technologies. CADTH. https://www.cadth.ca/guidelines-economic-evaluation-health-technologies-canada-0. Accessed May 3, 2020.
5. Putnick DL, Bornstein MH. Measurement invariance conventions and reporting: the state of the art and future directions for psychological research. *Dev Rev*. 2016;41:71–90.
6. Fayers P, Machin D. *Quality of Life: The Assessment, Analysis and Reporting of Patient-Reported Outcomes*. 3rd ed. Chichester, United Kingdom: John Wiley & Sons; 2016.
7. Later Life in the United Kingdom. Age UK. https://www.ageuk.org.uk/globalassets/age-uk/documents/reports-and-publications/later_life_uk_factsheet.pdf. Accessed December 15, 2019.
8. Proposed working definition of an older person in Africa for the MDS Project. World Health Organization. http://www.who.int/healthinfo/survey/ageingdefnolder/en/. Accessed October 16, 2018.
9. Milte CM, Walker R, Luszcz MA, Lancsar E, Kaambwa B, Ratcliffe J. How important is health status in defining quality of life for older people? An exploratory study of the views of older South Australians. *Appl Health Econ Health Policy*. 2014;12(1):73–84.
10. Ratcliffe J, Lancsar E, Flint T, et al. Does one size fit all? Assessing the preferences of older and younger people for attributes of quality of life. *Qual Life Res*. 2017;26(2):299–309.
11. General lifestyle survey: 2011. Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/compendium/generallifestylesurvey/2013-03-07. Accessed October 15, 2018.
12. Hospital episode statistics: admitted patient care, England - 2014-15. Health & Social Care Information Centre. https://files.digital.nhs.uk/publicationimport/pub19xxx/pub19124/hosp-epis-stat-admi-summ-rep-2014-15-rep.pdf. Accessed October 15, 2018.
13. Life expectancy at birth and at age 65 by local areas in the United Kingdom: 2006-08 to 2010-12. Office of National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/bulletins/lifeexpectancyatbirthandatage65bylocalareasintheunitedkingdom/2014-04-16. Accessed October 15, 2018.
14. Disability-free life expectancy (DFLE) and life expectancy (LE): at age 65 by region, England. Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/datasets/disabilityfreelifeexpectancydfleandlifeexpectancyleatage65byregionengland. Accessed October 14, 2018.
15. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011;20(10):1727–1736.
16. Devlin N, Parkin D, Janssen B. *Methods for Analysing and Reporting EQ-5D Data*. Berlin, Germany: Springer; 2020.
17. Mukuria C, Rowen D, Peasgood T, Brazier J. An Empirical Comparison of Well-Being Measures Used in UK. Policy Research Unit in Economic Evaluation of Health and Care Interventions (EEPRU), White Rose. https://eprints.whiterose.ac.uk/99499/1/EEPRU%20Report%20-%20Empirical%20comparison%20of%20well-being%20measures%20version%20interim%20report.pdf. Accessed February 15, 2020.

18. Scott NW, Fayers PM, Aaronson NK, et al. Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. *Health Qual Life Outcomes*. 2010;8:81.
19. Chang CH, Reeve BB. Item response theory and its applications to patient-reported outcomes measurement. *Eval Health Prof*. 2005;28(3):264–282.
20. Hays R, Morales L, Reise S. Item Response Theory and Health Outcomes Measurement in the 21st Century. *Med Care*. 2000;38(suppl 9):II28–II42.
21. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care*. 2007;45(5 suppl 1):S22–S31.
22. Kaiser H. The application of electronic computers to factor analysis. *Educ Psychol Meas*. 1960;20(1):141–151.
23. Cattell R. The scree test for the number of factors. *Multivariate Behav Res*. 1966;1(2):245–276.
24. Yu C. *Evaluating Cutoff Criteria of Model Fit Indices for Latent Variable Models With Binary and Continuous Outcomes*. Los Angeles, CA: University of California; 2002.
25. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual Life Res*. 2007;16(suppl 1):5–18.
26. Muthén LK, Muthén BO. Mplus user's guide. 7th ed. Muthén & Muthén. https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf. Accessed September 17, 2017.
27. Jones R, Tommet D, Ramirez M, Jensen R, Teresi J. Differential item functioning in Patient Reported Outcome Measurement Information system (PROMIS) Physical Functioning short forms: analyses across ethnically diverse groups. *Psychol Test Assess Model*. 2016;58(2):371–402.
28. Milfont T, Fischer R. Testing measurement invariance across groups: applications in cross-cultural research. *Int J Psychol Res*. 2010;3(1):111–121.
29. Vandenberg R, Lance C. A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ Res Methods*. 2000;2(1):4–69.
30. Teresi JA, Ocepek-Welikson K, Kleinman M, et al. Evaluating measurement equivalence using the item response theory log–likelihood ratio (IRTLR) method to assess differential item functioning (DIF): applications (with illustrations) to measures of physical functioning ability and general distress. *Qual Life Res*. 2007;16(suppl 1):43–68.
31. Meade AW. A taxonomy of effect size measures for the differential functioning of items and scales. *J Appl Psychol*. 2010;95(4):728–743.
32. Zumbo B. *Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense; 1999.
33. Bjorner JB, Kosinski M, Ware JE. Calibration of an item pool for assessing the burden of headaches: an application of item response theory to the headache impact test (HIT). *Qual Life Res*. 2003;12(8):913–933.
34. Keetharuth AD, Bjorner JB, Barkham M, Browne J, Croudace T, Brazier J. An item response theory analysis of an item pool for the recovering quality of life (ReQoL) measure. *Qual Life Res*. 2021;30(1):267–276.
35. Martin M, Blaisdell B, Kwong JW, Bjorner JB. The Short-Form Headache Impact Test (HIT-6) was psychometrically equivalent in nine languages. *J Clin Epidemiol*. 2004;57(12):1271–1278.
36. *SPSS. Statistics for Windows* [computer program]. Version 27.0. Armonk, NY: IBM Corp; 2020.
37. Smith AB, Cocks K, Parry D, Taylor M. A differential item functioning analysis of the EQ-5D in cancer. *Value Health*. 2016;19(8):1063–1067.
38. Smith AB. Differential item functioning and the Eq-5d: evidence from the UK Hospital episode statistics. *Value Health*. 2014;17(7):A514.
39. Fleishman JA, Lawrence WF. Demographic variation in SF-12 scores: true differences or differential item functioning? *Med Care*. 2003;41(7 suppl):III75–III86.
40. Lix LM, Wu X, Hopman W, et al. Differential item functioning in the SF-36 physical functioning and mental health sub-scales: a population-based investigation in the Canadian multicentre osteoporosis study. *PLoS One*. 2016;11(3):e0151519.
41. Yu YF, Yu AP, Ahn J. Investigating differential item functioning by chronic diseases in the SF-36 health survey: a latent trait analysis using MIMIC models. *Med Care*. 2007;45(9):851–859.
42. Penton H, Dayson C, Hulme C, Young T. A think aloud content validation of the EQ-5D-5L, SF-12v2, WEMWBS and ONS-4 in measuring the quality of life of older adults. Vol. 2020. In press.
43. Hulme C, Long AF, Kneafsey R, Reid G. Using the EQ-5D to assess health-related quality of life in older people. *Age Ageing*. 2004;33(5):504–507.
44. Mallinson S. Listening to respondents: a qualitative assessment of the Short-Form 36 Health Status Questionnaire. *Soc Sci Med*. 2002;54(1):11–21.
45. Moser DK, Heo S, Lee KS, et al. 'It could be worse… lot's worse!' Why health-related quality of life is better in older compared with younger individuals with heart failure. *Age Ageing*. 2013;42(5):626–632.
46. Ubel P, Jankovic A, Smith D, Langa K, Fagerlin A. What is perfect health to an 85-year-old?: evidence for scale recalibration in subjective health ratings. *Med Care*. 2005;43(10):1054–1057.
47. Tay L, Meade A, Cao M. An overview and practical guide to IRT measurement equivilence analysis. *Organ Res Methods*. 2015;18(1):3–46.
48. Teresi JA, Ocepek-Welikson K, Kleinman M, et al. Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): an item response theory approach. *Psychol Sci Q*. 2009;51(2):148–180.
49. Bilbao A, Martín-Fernández J, García-Pérez L, et al. Psychometric properties of the EQ-5D-5L in patients with major depression: factor analysis and Rasch analysis. *J Ment Health*. 2021:1–11.
50. Stochl J, Croudace T, Perez J, et al. Usefulness of EQ-5D for evaluation of health-related quality of life in young adults with first-episode psychosis. *Qual Life Res*. 2013;22(5):1055–1063.
51. Finch AP, Brazier JE, Mukuria C, Bjorner JB. An exploratory study on using principal-component analysis and confirmatory factor analysis to identify bolt-on dimensions: the EQ-5D case study. *Value Health*. 2017;20(10):1362–1375.
52. Keeley T, Al-Janabi H, Fletcher K, McManus R, Mant J, Coast J. An assessment of the validity and responsiveness of the ICECAP-O in a multicentre randomized controlled trial of blood pressure management. *Value Health*. 2016;19(3):A93.
53. Davis JC, Liu-Ambrose T, Richardson CG, Bryan S. A comparison of the ICECAP-O with EQ-5D in a falls prevention clinical setting: are they complements or substitutes? *Qual Life Res*. 2013;22(5):969–977.