HEDG
HEALTH, ECONOMETRICS AND DATA GROUP
THE UNIVERSITY of York

WP 16/14

Differential item functioning in the EQ-5D: An exploratory analysis using anchoring vignettes

Rachel Knott, Paula Lorgelly, Nicole Black & Bruce Hollingsworth

August 2016

http://www.york.ac.uk/economics/postgrad/herc/hedg/wps/

Differential item functioning in the EQ-5D: An exploratory analysis using anchoring vignettes

Rachel Knott[1*], Paula Lorgelly[1,2], Nicole Black[1], Bruce Hollingsworth[3]

[1] Centre for Health Economics, Monash University
[2] Office of Health Economics, London
[3] Division of Health Research, Lancaster University
[*] *Corresponding author*

Abstract

Inter-group comparisons using the EQ-5D, or any self-reported measure of health, rely on the measure being an accurate reflection of the true health of the groups or individuals concerned. However, responses to questions on subjective scales, such as those used in the EQ-5D, will be inaccurate if groups of individuals systematically differ in their use of the response categories, a phenomenon known as *differential item functioning* (DIF). This paper reports on an exploratory analysis involving the use of anchoring vignettes to identify differential item functioning (DIF) in the EQ-5D-5L. We demonstrate that using vignettes to appropriately identify DIF in EQ-5D reporting is possible, at least in certain age groups. We find that the EQ-5D is indeed subject to DIF, and that failure to account for DIF can lead to conclusions that are misleading when using the instrument to compare health or quality of life across heterogeneous groups. For instance, when adjusting for DIF in a sample aged 55-65 years, we found that differences between the highest and lowest education groups doubled in value after adjusting for DIF, and increased from quantities that would not have had relevance in a clinical settings to ones that would (based on a suggested minimally important difference). Thus, our research provides evidence that the EQ-5D should be used with caution when comparing health or quality of life across heterogeneous groups. We also provide several important insights in terms of the identifying assumptions of response consistency and vignette equivalence.

Acknowledgements: This research was funded by an Australian Research Council Discovery Project Grant (DP110101426), a BankWest Curtin Economics Centre (BCEC) grant, and a Monash Faculty of Business and Economics grant.

Key words: Differential item functioning, Anchoring vignettes, EQ-5D, Response consistency, Vignette equivalence

JEL Classification: I10, C19, C49

1. Introduction

Categorical response scales (e.g. *excellent health* to *poor health*; *no problems* to *extreme problems*) used to measure self-reported health are integral components of decision-making across a range of health and medical research settings. They are used to assess effectiveness, inequalities and general health status; however, as these measures are by nature subjective, response categories can often mean different things to different people. Systematic differences in the ways that people use and interpret response categories can introduce bias when using self-reports to compare health or quality of life across heterogeneous patient or population groups. For example, people may rate their health differently, not only because their true or perceived health differs, but also because they interpret and use the response scales differently; thus seemingly important differences may actually be explained, at least in part, by differential use of response categories. This is a phenomenon known as *reporting heterogeneity*, *response-scale heterogeneity* or *differential item functioning* (DIF) (King et al., 2004). DIF has been shown to exist across a range of subject areas, including in other self-reported measures of health (Kapteyn et al., 2007, e.g. Bago D'Uva et al., 2008b, Grol-Prokopczyk et al., 2011), but has largely been overlooked in the case of the increasingly popular Patient Reported Outcome Measures (PROMs).

The most commonly used PROM is the EuroQol's EQ-5D (Brooks, 1996), which asks respondents to describe their health on five different "*dimensions*": mobility, self-care, usual activities, pain/discomfort and anxiety/depression (Devlin and Krabbe, 2013, Dolan et al., 2013). Typically, these responses are converted to a preference-based weighted summary score or index which reflects a health state utility value (where 0 is dead and 1 is full health). Researchers often combine these scores with length of life to obtain quality-adjusted life years (QALYs) for use in economic evaluations of heath technology assessments (HTAs). More recently the EQ-5D has also been used as a measure of population health status and is included in a number of population health surveys globally (Euroqol Group, 2014, Burström et al., 2001).

In both settings, the instrument is often used to compare health related quality of life (HRQoL) across patient or population groups. For example, in population health research it has been used to compare HRQoL across groups according to diagnosed or self-reported health conditions (Lubetkin et al., 2005, Ko and Coons, 2006, Sullivan and Ghushchyan, 2006, Devlin et al., 2010), behavioural risk factors (Søltoft et al., 2009, Maheswaran et al., 2012), and socio-demographic characteristics such as age, gender and socio-economic status (Kind et al., 1998, Burström et al., 2001, Lubetkin et al., 2005, Luo et al., 2005, Sun et al., 2011). In the case of economic evaluations, sub-group analyses are often used to identify cost-effective populations, e.g. according to age groups (Prosser et al., 2000). On a broader scale, when used in submissions to reimbursement agencies to inform decisions about which tests, treatments and health care

interventions to fund, PROMs (in the form of QALYs) can be implicitly compared across all demographic groups for alternate interventions. If inter-group comparisons using the EQ-5D are affected by DIF, it may bring into question any perceived findings, for instance, that health inequalities exist (in population health research), or that health care interventions are cost-effective (in health economic evaluations).

Two previous studies have found evidence of DIF in EQ-5D reporting across countries in the case of the original EQ-5D-3L (Salomon et al., 2011, Whynes et al., 2013).[1] While it has not been previously tested, it is also likely that DIF in the EQ-5D extends to subgroups within countries; for example, according to groups divided by gender, age, or socioeconomic status. To address DIF, Salomon et al. (2011) and Whynes et al. (2013) made use of more objective measures, such as detailed health instruments and clinical measures, to separate differences in health from differences in reporting styles. This approach is valid so long as the objective measures adequately capture variation in underlying latent health for each of the dimensions (which can be difficult to achieve in practice); if not the result will be confounded by unobserved influences (Salomon et al., 2011).

Another method which has been described as "the most promising" approach for detecting DIF is the use of a survey tool known as the anchoring vignette approach (Murray et al., 2002 p.429). The method involves the inclusion of at least one, but typically several, health descriptions of hypothetical individuals (the vignettes) that respondents are asked to rate in addition to rating their own health using the same subjective ordered categories. Provided that two key identifying assumptions hold, namely response consistency (RC) and vignette equivalence (VE), these ratings can reveal what the response categories truly mean for respondents, and can therefore be used to identify and adjust for DIF. The approach has been used in a number of applications including political efficacy (King et al., 2004), job, income and life satisfaction (Kristensen and Johansson, 2008, Kapteyn et al., 2013, Angelini et al., 2014, Bertoni, 2015) and general and specific dimensions of health (Kapteyn et al., 2007, Bago D'Uva et al., 2008b, Grol-Prokopczyk et al., 2011, Bago d'Uva et al., 2011, Molina, 2016).

Anchoring vignettes provide a convenient alternative to the collection of gold standard objective measures, which can be expensive and inconvenient to collect particularly in self-completion style questionnaires (Grol-Prokopczyk et al., 2015). Moreover, many measures of interest, such as levels of pain or usual activities (both dimensions of the EQ-5D) cannot be measured objectively. The anchoring vignette approach could therefore potentially serve as a viable means for identifying and adjusting for DIF in the EQ-5D. However, the appropriate use of

[1] The new version, the EQ-5D-5L, has been expanded to five "levels" or response options – no problems, slight problems, moderate problems, severe problems and extreme problems – which could further increase the potential for DIF.

vignettes relies on the assumptions of RC and VE, which, are proving difficult to achieve in practice. Earlier studies that adopted informal and often minimal approaches to test these assumptions tended to endorse the validity of the anchoring vignette approach (Kristensen and Johansson, 2008, Grol-Prokopczyk et al., 2011); a number of more recent studies that have applied newly developed, rigorous econometric tests have called into question whether the assumptions in fact hold (Bago d'Uva et al., 2011, Peracchi and Rossetti, 2013, Grol-Prokopczyk et al., 2015).

This study investigates whether the anchoring vignette approach can appropriately be used to identify DIF in the EQ-5D-5L. The paper builds on the work of Au and Lorgelly (2014) who developed EQ-5D-5L-specific anchoring vignettes and qualitatively examined their performance in relation to the assumption of RC in a smaller pilot study. We extend this analysis to a larger sample of respondents from the general population to formally test the appropriateness of using anchoring vignettes for the EQ-5D by employing recently developed methods of Bago d'Uva et al. (2011), which have been referred to in the literature as 'strong' tests for RC and VE (Grol-Prokopczyk et al., 2015). Our results for RC are very promising; the assumption holds for all five dimensions for our representative sample. However, we find that VE holds only for a subgroup of older respondents (age 55-65). We focus on this subsample, for which both assumptions hold, to assess whether the EQ-5D is indeed subject to heterogeneity in reporting styles. Finally, we examine the impact of DIF-bias on inter-group comparisons using EQ-5D indices. We do this by estimating EQ-5D scores that would have prevailed, had respondents evaluated their health on common response scales (i.e. common across all respondents), and comparing these to unadjusted measures.

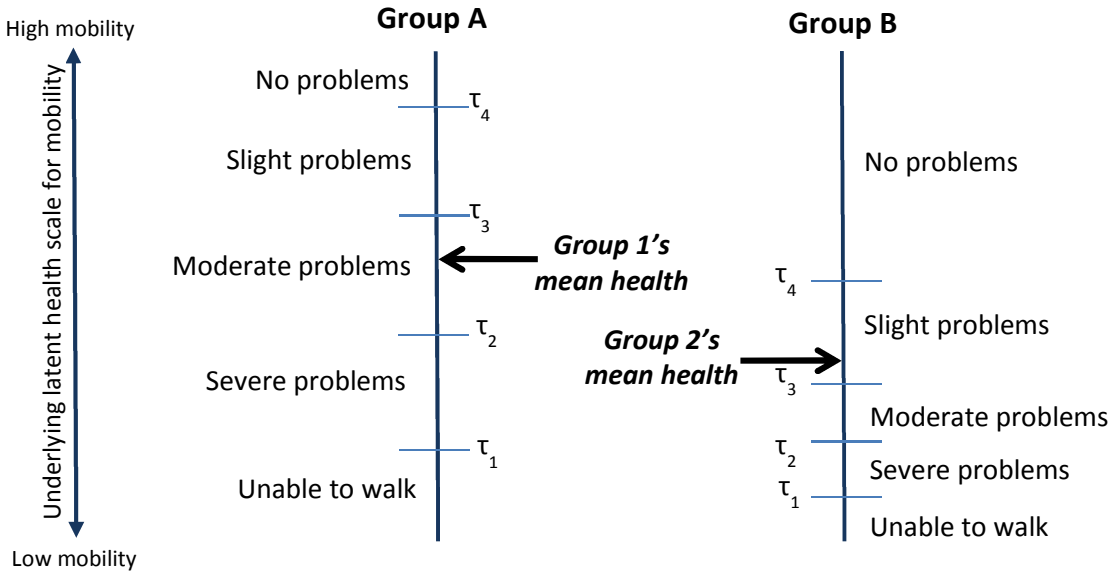
The paper proceeds as follows. The next section discusses the intuition of the anchoring vignette approach and the assumptions of RC and VE. Section 3 outlines the methodology we use to test the identifying assumptions, and to test and adjust for DIF in the EQ-5D-5L. Section 4 describes the vignettes and survey design in detail, while section 5 presents our results and robustness checks. The final section concludes.

2. Differential item functioning and anchoring vignettes

DIF is illustrated in Figure 1. For each health dimension, there is a latent scale that is unobserved, which is represented by the vertical line. We take the example of the single dimension for mobility, which is the first EQ-5D-5L dimension, and compare hypothetical response categories for two groups of people: Group A and Group B. Individuals in both groups are asked to self-report their own level of mobility using the five response options: no problems in walking about; slight problems in walking about; moderate problems in walking about; severe problems in walking about; and unable to walk about. How the average individual in each group divides the underlying latent scale into the five levels or response categories – or alternatively, the placement of the inter-

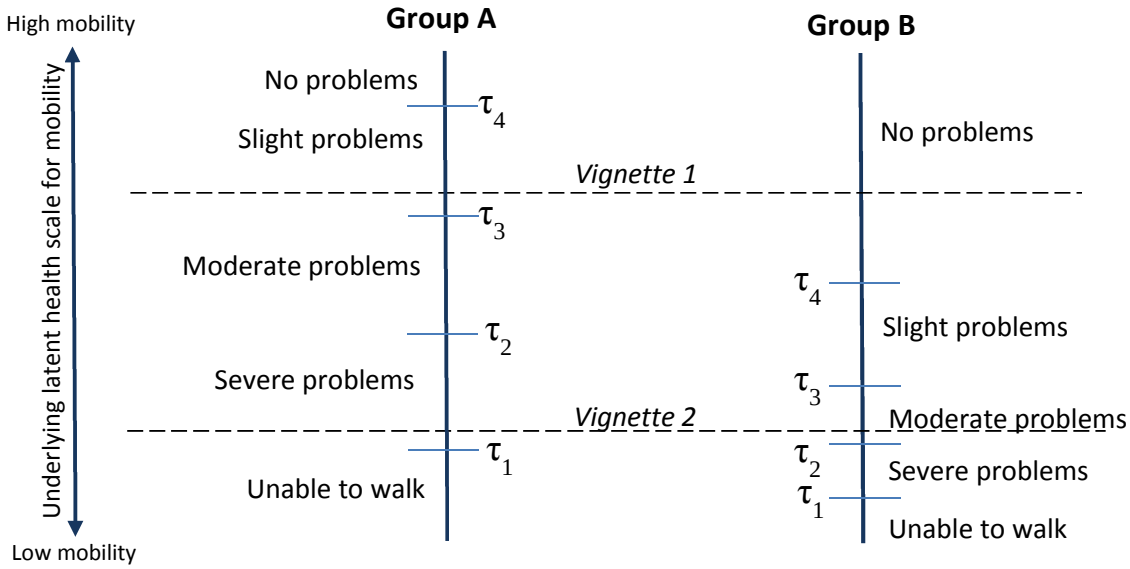
category thresholds – is represented by τ_1 to τ_4 (i.e. the first to fourth thresholds). DIF is portrayed in the figure by variation in the placement of the thresholds across the two groups. Despite Group A having a higher mean level of underlying latent mobility compared to Group B, as evidenced by the bold arrow being placed higher up the scale, Group A reports moderate problems on average while Group B reports only slight problems. In this example Group B is more health optimistic compared to Group A, however this is typically not evident to researchers, who would incorrectly infer that Group B has a higher level of mobility.

Figure 1 – Example of DIF



In order to obtain any meaningful comparison between the health of Groups A and B it is essential to adjust for DIF (Murray et al., 2002). Anchoring vignettes can be used to do this adjustment (King et al., 2004), where a vignette is a brief description of a health state of a hypothetical individual. Typically, a series of vignettes are presented for each health construct of interest, at varying levels of severity. Suppose we have two vignettes, where the person in vignette 1 is described to have fewer mobility problems compared to the person described in vignette 2. How Groups A and B rate the health of the vignettes on average is illustrated in Figure 2 (where the fixed health of each vignette is represented by the dotted horizontal lines). Group B's relative health optimism is evident upon consideration of vignette assessments: Group B's ratings are more favourable than Group A's ratings for both vignettes (i.e. mean ratings for the two vignettes are slight problems (vignette 1) and severe problems (vignette 2) for Group A, and no problems (vignette 1) and moderate problems (vignette 2) for Group B). Vignettes can therefore help to identify differential reporting behaviour.

Figure 2 – Logic underlying anchoring vignettes to locate respondent thresholds



2.1. Response consistency and vignette equivalence

The anchoring vignette approach rests on the identifying assumptions of response consistency (RC) and vignette equivalence (VE). RC is the assumption that respondents rate the health of the hypothetical people described in the vignettes using the same underlying scale that they use to rate their own health. RC would be violated if, for example, respondents rated the health described by the vignettes either more or less harshly than they did their own health. If RC fails to hold, the thresholds identified by the anchoring vignettes would not be the same as those that individuals use to identify their own health, thus DIF cannot be adequately identified. VE holds if all respondents interpret the health states described by the vignettes in the same way and on the same uni-dimensional scale, aside from random error (King et al., 2004), and is represented in our example by the fixed horizontal lines of Figure 2. This assumption is essential for the vignettes to act as an anchor; such that all systematic variations between vignette ratings and individual characteristics can be attributed to DIF (Bago d'Uva et al., 2011). We return to these assumptions in Section 4.

3. Data and vignette design

The analysis is based on two online surveys. An initial survey was conducted in April 2014 and involved 1,007 respondents. A second survey, aimed at gaining more data points, was then carried out between August and September of 2015 and involved an additional 3,293 respondents, yielding a total sample size of 4,300. The sampling strategy targeted a representative sample of Australians aged 18 to 65 (in terms of gender-age-State of residence splits) who were recruited via a survey panel company. The surveys collected information from respondents on standard socio-economic and demographic variables, self-reports of the presence of health conditions (e.g. diabetes or

cancer), self-reports of their HRQoL using the EQ-5D-5L, and the anchoring vignettes (described below). Additionally, the initial survey contained a range of supplementary health questions related to each dimension of the EQ-5D[2], which were included to proxy 'objective' health in the tests for RC described in Section 4.2.1. For the mobility dimension we adapted the detailed 'objective' set of mobility questions used by Kapteyn et al. (2011) and constructed a similar question set for the self-care and usual activities dimensions (available on request). The Short-Form McGill Pain Questionnaire (SF-MP) was utilized as an objective measure for pain; and the Kessler index (K-10) as a measure to gauge anxiety/depression. Questions about the presence of health conditions more generally were also included as objective measures for all dimensions (in addition to those mentioned above), such as the presence of diabetes, osteoporosis or cancer. Ethics approval was obtained by the <removed to maintain author anonymity>.

3.1. EQ-5D vignettes

The vignettes used in this study were based on those developed by Au and Lorgelly (2014), which investigated the assumption of RC in the EQ-5D-5L using face-to-face opened-ended interviews as well as a series of questions administered in an online survey.[3] Two vignettes were shown, representing differing levels of health, which provided complete health state descriptions covering all EQ-5D dimensions (vignettes are presented in Appendix A). The vignettes were adapted slightly from the vignettes developed in Au and Lorgelly (2014) according to various findings of their qualitative analysis. For instance, we removed any mention of (what could be considered) age-related health conditions such as arthritis, because it was shown that younger respondents found it difficult to imagine themselves in such health states, thus jeopardising RC (respondents are thought to be more likely to respond consistently if they think of the vignette persons as similar to themselves).

Before the vignettes appeared in the survey we gave succinct instructions which were found to be effective at enhancing the potential for RC in the qualitative study (Au and Lorgelly, 2014). Respondents were asked to imagine themselves in the health state of the individuals in the hypothetical scenarios when rating the vignettes. They were also asked to imagine the hypothetical persons as having the same age and background as themselves (Jürges and Winter, 2013).

As recommended by King et al. (2004), the questionnaires were gender specific so that the names of the hypothetical people in the vignettes were the same sex as the respondent (this is also thought to encourage respondents to think of the vignette persons as similar to themselves (Hopkins

[2] These additional questions were not included in the second survey due to budgetary limitations.

[3] Au and Lorgelly (2014) developed two types of vignettes for the EQ-5D, namely shorter vignettes specific to each EQ-5D dimension (Version A); and longer, holistic vignettes describing overall states of health (Version B, which are essentially Version A vignettes combined at each level of severity). Qualitative analyses of the pilot study found response consistency was more likely to hold for Version B, therefore the current analyses uses Version B vignettes.

and King, 2010)), and names were selected from a list of the most popular names since the 1940s that appeared across several decades (Jürges and Winter, 2011). The order of the vignettes in terms of severity was randomised across respondents. Vignettes were placed after the EQ-5D-5L self-assessment in both surveys, and in the initial survey the additional health questions were placed after the EQ-5D-5L and vignettes – this was done so that the responses to the EQ-5D-5L were not influenced by the vignettes or any further health questions.

3.2. Selection of variables to identify DIF and analysis sample

We are particularly interested in assessing the impacts of DIF across age, gender, education and country of birth, all of which have been shown to affect DIF in various health dimensions (e.g. Bago d'Uva et al., 2008a, Bago D'Uva et al., 2008b, Grol-Prokopczyk et al., 2011, Molina, 2016). Age is considered in terms of four categories: 20 to 34; 35 to 44; 45 to 54 and 55 to 65 at the time of survey.[4] Education was represented by three dummy variables representing highest educational attainment: year 12 schooling or less (herein referred to as *low*), trade certificate or diploma (*medium*) and university degree (*high*). Country of birth was divided into four categories: Australia; other English speaking countries; Asian countries; and other non-English speaking countries (referred to as *other*). Marital and employment status were also included as they too have been found to influence reporting styles in other analyses (Kapteyn et al., 2013). Marital status was grouped into three categories: married or de facto relationship; divorced, separated or widowed; and never married. Employment status was also represented by three dummy variables: employed; unemployed; and retired or not in the labour force (NILF).

As we are interested in comparing EQ-5D scores across levels of education, we removed respondents who were aged less than 20 years from the analysis as they may not yet have finished their studies (120 respondents), as well as those aged over 20 that indicated they were still studying at the time of the survey (85). This left us with a pooled sample size of 4,095 (973 from the initial survey and 3,122 from the second). Respondent characteristics of this sample, including average vignette ratings, are provided in Table B.1 of Appendix B.

4. Econometric approach

4.1. Formal testing and adjustments for DIF

To formally examine DIF we use the hierarchical ordered probit (HOPIT) model with anchoring vignettes introduced by King et al. (2004). The HOPIT is an extension of the ordered probit (OP) model which allows for variation in the inter-category thresholds by modelling them as a function

[4] Convergence issues were experienced in tests for RC and VE when more age dummies were included.

of covariates (cf. the OP model which holds constant the location of the thresholds across respondents).

The likelihood function of the HOPIT is made up of a self-assessment component and a vignette component. For the self-assessment component we assume that respondent i's observed response for a particular dimension of the EQ-5D-5L, y_i, is associated with an underlying latent variable Y_i^* for the particular dimension of interest (e.g. mobility). The latent health variable is characterized by the relationship

$$Y_i^* = X_i\beta + \varepsilon_i, \quad (1)$$

where X_i is a set of covariates and ε_i follows a standard Normal distribution with $\varepsilon_i \sim N(0, \sigma^2)$ (scale and location restrictions are imposed that constrain the constant in X_i to zero and σ^2 to 1). y_i is observed according to the mechanism:

$$y_i = k \; if \; \tau_i^{k-1} < Y_i^* \leq \tau_i^k$$
$$-\infty = \tau_i^0 < \tau_i^1 < \tau_i^K = \infty. \quad (2)$$

Here τ_i^k represents the inter-category threshold for the k^{th} response. The thresholds are a function of covariates Z_i such that

$$\tau_i^1 = \gamma^1 Z_i$$
$$\tau_i^k = \tau_i^{k-1} + \exp(\gamma^k Z_i) \text{ for k=2,…5.} \quad (3)$$

In our case Z_i and X_i are equivalent. For the vignettes component we assume that the underlying latent health described by vignette j is fixed at α_j (representing the assumption of VE). Respondent i's perceived health state for vignette j is given by

$$V_{ij}^* = \alpha_j + \xi_{ij} \quad (4)$$

(where ξ_{ij} is random measurement error with $\xi_{ij} \sim N(0, \sigma_v^2)$),[5] which is rated on the ordered categorical scale according to

$$V_{ij} = k \; if \; v_i^{k-1} < V_{ij}^* \leq v_i^k$$
$$-\infty = v_i^0 < v_i^1 < v_i^K = \infty, \quad (5)$$

where v_i^k represent inter-category thresholds for the vignette responses, such that

$$v_i^1 = {\gamma_v}^1 Z_i$$
$$v_i^k = v_i^{k-1} + \exp({\gamma_v}^k Z_i) \text{ for k=2,…5.} \quad (6)$$

Under the assumption of RC, we have

$$\gamma^k = {\gamma_v}^k, \quad (7)$$

and therefore, the responses to the vignettes can be used to identify γ^k in the estimation procedure. The model is estimated by way of maximum likelihood, where the likelihood equation is a function

[5] While it is possible to estimate the variance term σ_v^2 for each vignette, it is typically assumed constant across vignettes for simplicity.

of both the likelihood of the self-assessment component, $L_s(\beta, \gamma|y)$, and the likelihood of the vignettes component, $L_v(\theta, \gamma|v)$, such that

$$L(\beta, \theta, \gamma|y, v) = L_s(\beta, \gamma|y) \times L_v(\theta, \gamma|v). \quad (8)$$

We estimate five separate HOPIT models for each dimension of the EQ-5D-5L. The presence of DIF is formally tested using likelihood ratio (LR) tests that restrict the threshold covariates to zero, i.e. $\gamma^k = 0$ for $k = 1, \dots, K-1$.

The impacts of DIF are then assessed by estimating, for each individual, the EQ-5D score that would have prevailed had all respondents evaluated their health states using the same underlying response scales. DIF-adjusted indices are estimated by first conducting counterfactual simulations to obtain DIF-adjusted outcomes for each of the EQ-5D-5L dimensions. This is done by simulating the distribution of Y_i^* using the estimated parameters from the mean function of the HOPIT models (i.e. $\hat{\beta}$) and the characteristics of each individual, X_i. For each dimension the simulated latent index is then converted back to the EQ-5D-5L response categories or levels by applying the predicted thresholds, $\widehat{\tau^k}$, at sample means, $\bar{X}$ (Kapteyn et al., 2013, Angelini et al., 2014, Bertoni, 2015). EQ-5D-5L tariffs (valued using an Australian discrete choice experiment (Norman et al., 2013)) are then applied to the reported and DIF-adjusted health profiles. The overall impact of DIF on inter-personal comparisons of health is determined by comparing unadjusted and DIF-adjusted summary indices across sub-groups according to the characteristics described in section 3.2. A bootstrap procedure is conducted to determine whether observed differences across indices are statistically significant.

4.2. Testing for response consistency and vignette equivalence

4.2.1. Response consistency

We adopt tests for RC and VE developed by Bago d'Uva et al. (2011). For RC, the authors note that if objective measures (H^O) are available, such that

$$h(Y_i^*|H^O, X) = h(Y_i^*|H^O), \quad (9)$$

where $h(.)$ is a density function for latent health, then the response scales that respondents use to report their own health can be identified by a model defined by

$$Y_i^* = g\left(H_i^O\right) + \eta_i, \quad (10)$$

combined with equations 2 and 3 above, where $g(.)$ is the same across individuals and η_i is random error. By jointly estimating this model with a model defined by equations 4, 5 and 6, RC – the equality of equation 7 – can be formally tested.

As noted by the authors, this test may be too restrictive in that it could fail if the objective measures do not adequately capture variations in health, or if VE is violated. Bago d'Uva et al. therefore propose a second, more robust test which does not rely on these assumptions, and states

that the distances between any two inter-category thresholds should be the same across the equation for health and vignettes, i.e.

$$\gamma^k - \gamma^{k-1} = {\gamma_v}^k - {\gamma_v}^{k-1}, k = 2, \dots, K-1 \quad (11)$$

which forms the null hypothesis. As we find VE does not hold for our sample (see below), and that our objective measures, being self-reported, may suffer their own shortcomings (DIF being one of them; the lack of objectivity being another), we adopt this second more robust test in our analysis.

4.2.2. Vignette equivalence

The VE assumption imposes a restriction that X must be excluded from equation 4; that is, there must not be systematic differences in the perceptions of the health states of the vignettes persons. Bago D'Uva and colleagues test the assumption by estimating a model that extends equation 4 to include interactions between individual characteristics and vignette severity for all but one vignette (if all vignettes are included the model is not identified); i.e.

$$V_{i1}^* = \alpha_1 + \upsilon_{i1}$$
$$V_{ij}^* = \alpha_j + \lambda_j X_i^- + \upsilon_{ij} \text{ for } j = 2, \dots J \quad (12)$$

where X^- is X without the constant term, λ_j is the vector of coefficients to be estimated, and $\upsilon_{ij} \sim N(0, \sigma_v^2)$. $\lambda_j \neq 0$ would suggest systematic differences in the perception of the j^{th} vignette relative to the first (the reference), which would violate VE. The null hypothesis is therefore

$$\lambda_j = 0 \text{ for } j = 2, \dots J \quad (13)$$

which is examined using a LR test that compares the restricted model given by equations 4, 5 and 6 with the unrestricted model of equations 12, 5 and 6.

5. Results

5.1. Assumption tests

Tests for RC were focused on the sample of 973 respondents that answered the 'objective' health questions (i.e. the initial survey). Convergence issues were experienced when conducting the test using all five outcome categories, presumably because of the relatively small sample size. Following Bago d'Uva et al. (2011), we therefore collapsed the five outcome categories to three by combining the categories for no problems with slight problems, and severe problems with extreme problems.

Results for the RC tests are presented at the top of Table 1, and pass for all five dimensions at the 5% level of significance. This result is distinct from a number of other studies that have considered formal tests for RC - often it is found that the assumption does not hold (Kapteyn et al., 2011, Bago d'Uva et al., 2011, Peracchi and Rossetti, 2013). For example, Bago d'Uva et al. (2011) find strong evidence against RC when considering mobility vignettes in the English Longitudinal

Study of Ageing (ELSA) survey. Our contrasting (favourable) result may be due, in part, to the attention exerted in the survey design aimed at improving the likelihood of achieving RC.

Table 1 – Tests of vignette assumptions

	Degrees of freedom	χ^2 test statistic	*p*-value
Response consistency			
Mobility	13	15.12	0.300
Self-care	13	18.31	0.146
Usual activities	13	8.14	0.835
Pain/discomfort	13	18.86	0.127
Anxiety/depression	13	19.44	0.110
Vignette equivalence			
Mobility	13	100.06	<0.001
Self-care	13	178.69	<0.001
Usual activities	13	170.03	<0.001
Pain/discomfort	13	241.63	<0.001
Anxiety/depression	13	172.44	<0.001

Note: The test for VE was conducted on all respondents of the analysis sample (n = 4,095); while the RC test focussed on the subsample of respondents that answered the 'objective' health questions (n = 973).

Results for the VE test are presented at the bottom of Table 1. VE is rejected for all EQ-5D dimensions, suggesting that systematic differences occur across the perception of the health states described by vignettes. It is conceivable that this too could be an artefact of the survey design, albeit a negative one. In particular, the attempted avoidance of age-related issues and the preliminary instructions asking respondents to imagine that the vignettes were of a similar age and background to themselves, may have opened up the potential for variations in the perceptions of the vignette-described health states, thus jeopardizing VE. RC may therefore have come at the expense of VE.

While VE was not found to hold for our sample at large, it may be that the assumption holds within certain groups of the population. Specifically, since respondents were asked to consider the persons in the vignettes to be of a similar age as themselves, VE may hold for respondents of similar ages. To examine this conjecture, we repeated the VE tests in each of the age groups described in section 3.2. The results for these tests are presented in Table 2. While the assumptions did not hold for the 20-34, 35-44, and 45-55 year age groups, they were found to hold for the 55-65 years group for all dimensions except anxiety/depression (which we return to below). Why VE held only for the older age group and not for the younger age groups we are unable to say, but it may have to do with an inability of (some) younger individuals to conceptualize vignettes describing

situations of unfavourable health, leading to greater variations in the interpretation of these vignettes. Older individuals, on the other hand, may be in a better position to understand these states of health, as they are more likely to have observed them either through personal experience or through the experience of their peers. If this were the case, it may be that vignette 1 – our least severe health description - has a greater potential for VE than vignette 2 amongst younger age groups. Unfortunately however we are not able to test this hypothesis, as the test for VE requires more than one vignette.

Table 2 – Vignette equivalence – alternate age groups

	Degrees of freedom	χ^2 test statistic	*p*-value
Age 20-34			
Mobility	8	21.785	0.005
Self-care	8	65.791	<0.001
Usual activities	8	54.208	<0.001
Pain/discomfort	8	68.995	<0.001
Anxiety/depression	8	38.895	<0.001
Age 35-44			
Mobility	8	28.017	<0.001
Self-care	8	75.826	<0.001
Usual activities	8	56.664	<0.001
Pain/discomfort	8	79.472	<0.001
Anxiety/depression	8	45.601	<0.001
Age 45-54			
Mobility	8	67.563	<0.001
Self-care	8	110.842	<0.001
Usual activities	8	93.543	<0.001
Pain/discomfort	8	129.923	<0.001
Anxiety/depression	8	82.278	<0.001
Age 55-65			
Mobility	8	8.296	0.600
Self-care	8	9.427	0.492
Usual activities	8	11.675	0.307
Pain/discomfort	8	15.076	0.129
Anxiety/depression	8	24.061	0.007

Note: Age 20-23: n=834; Age35-44: n = 1476; Age 45-54: n=871; Age 55-65: n=914

As mentioned, VE did not hold for the anxiety/depression dimension. A closer inspection revealed that the reason for failure was the coefficient for females (in equation 12), which was negative and statistically significant (*p*-value = 0.002), suggesting that female respondents interpret vignette 2 as being closer to the reference category (vignette 1) than do males, in terms of

anxiety/depression. Kapteyn et al. (2013) also identified a violation of VE in their assessment of income satisfaction vignettes; however for a number of covariates, not just one. Overall the VE violation in their analysis was not found to bias their overall result. Making the same assumption, we progress our analysis on the sample aged 55 to 65 (N =914), for which we can be reasonably certain that the vignettes are adequately identifying DIF (see Table B.1 of Appendix B for characteristics of this sample). In robustness checks below we revisit the violation of VE across gender in the dimension of anxiety/depression.

5.2. Identification of DIF

LR tests for DIF are presented in Table 3 for the sample aged 55 to 65 years; that is, the sample among which we can be reasonably certain that the anchoring vignettes are appropriately identifying DIF. The null hypothesis of reporting homogeneity is rejected for all five dimensions at the 5% level, suggesting that DIF is present in the sample of 55-65 year olds. An inspection of the parameters in the threshold equations (Table 4 and Table B2 in Appendix B) suggests that the nature of heterogeneity in the use of the response scales varies across dimensions (for ease of interpretation the order of response categories have been reversed and rate from extreme limitations/unable (category 1) to no limitations (category 5)). For instance, education appears to affect reporting behaviour – at least to some degree – across all dimensions, while gender does not significantly influence reporting behaviour for self-care or usual activities.

Table 3 – Tests for reporting homogeneity for sample aged 55-65 years

	Mobility	Self-care	Usual activities	Pain/ discomfort	Anxiety/ depression
LR test statistic	94.82	57.71	64.73	74.89	74.57
p-value	<0.001	0.043	0.008	0.001	0.001
Degrees of freedom	40	40	40	40	40

Note: n=914

Table 4 – HOPIT estimates for first threshold

	Mobility	Self care	Usual activities	Pain/ Discomfort	Anxiety/ Depression
Female	-0.165*	-0.005	0.059	0.131***	0.035
	(0.087)	(0.052)	(0.046)	(0.050)	(0.047)
Education (base category low)					
Medium	-0.128	-0.088	0.014	-0.109*	0.047
	(0.095)	(0.061)	(0.054)	(0.057)	(0.055)
High	-0.251**	-0.168**	-0.073	-0.142**	-0.03
	(0.107)	(0.067)	(0.057)	(0.061)	(0.058)
Country of Birth (ref. Australia)					
Oth English speaking	0.099	0.125	-0.097	0.188**	0.119
	(0.160)	(0.095)	(0.094)	(0.089)	(0.088)
Asia	0.168	0.037	0.025	0.055	0.02
	(0.105)	(0.073)	(0.065)	(0.070)	(0.066)
Other	0.399**	0.159	0.142	0.201	0.118
	(0.179)	(0.133)	(0.121)	(0.126)	(0.123)
Marital status (ref. never married)					
Married/de facto	-0.335***	-0.165**	-0.005	-0.063	0.008
	(0.103)	(0.074)	(0.070)	(0.074)	(0.073)
Divorced/widowed	-0.259**	-0.123	0.066	-0.034	0.092
	(0.123)	(0.084)	(0.079)	(0.084)	(0.081)
Employment status (ref. NILF)					
Employed	-0.009	-0.032	-0.074	-0.044	-0.087*
	(0.084)	(0.053)	(0.048)	(0.051)	(0.048)
Unemployed	-0.333	-0.127	0.018	-0.023	-0.269**
	(0.265)	(0.128)	(0.102)	(0.113)	(0.120)
Constant	-1.517***	-1.452***	-1.578***	-1.649***	-1.518***
	(0.136)	(0.148)	(0.113)	(0.105)	(0.107)

* p<0.10, ** p<0.05, *** p<0.01.
Note: Sample aged 55-65 years (n=917); standard Errors in parentheses

In these tables, a positive coefficient signifies that respondents place the given threshold higher up the latent scale than the reference group, while a negative coefficient suggests the opposite. For example, focussing on the first threshold (Table 4), respondents who have obtained a university qualification (i.e. *high* education) place the threshold between "extreme problems" and "severe problems" lower down the latent scale than that of individuals whose highest education level is a high school degree or less (i.e. *low* education) for the dimensions of mobility, self-care and pain/discomfort.[6] This in turn means that the most educated in this sample are less likely to report the worst outcome (extreme problems) for these dimensions for a given level of health. This result, particularly for the first threshold, is corroborated by others including Grol-Prokopczyk et al.

[6] Note, however, that this effect is not always consistent across remaining thresholds – see Table B2 of Appendix B.

(2011) and Molina (2016). Similarly, people who are married or in de facto relationships are less likely to report extreme problems compared to individuals who never married, for mobility and self-care. On the other hand, people born in English speaking countries other than Australia are more likely to report extreme problems for pain or discomfort; thus heterogeneity in cultural backgrounds within a given country may also bring about DIF.

Covariates of the remaining thresholds are presented in an Appendix; although note that they are difficult to interpret directly since they depend on the locations of the preceding thresholds and are functions of exponentials. While not all covariates are significant (which could be due to sample size in the 55-65 year old subgroup), it is interesting that the magnitudes across covariates are largely consistent across the various dimensions. Also in Appendix B, for interested readers, we present the results for the mean function of each HOPIT model (Table B2). Following other authors (Kapteyn et al., 2007, Angelini et al., 2014, Bertoni, 2015), these results are presented alongside models that do not allow for DIF, i.e. OP models - the parameters of which can be directly compared to those of the HOPIT.[7]

5.3. DIF-adjustments in EQ-5D indices

Figure 3 illustrates the difference between the unadjusted and DIF-adjusted EQ-5D-5L scores, with associated 95% confidence intervals, for the sample aged 55 to 65; and Table 5 presents differences in mean indices across subgroups. The confidence intervals of Figure 3, and *p*-values of Table 5 are calculated using bootstrapped standard errors. The mean index for individuals aged 55 to 65 increases from 0.729 (unadjusted) to 0.806 (DIF-adjusted) where the upward adjustment reflects the movement of respondents away from the most limiting categories which impose the highest penalties to EQ-5D-5L scores (see Appendix B, Table B4, for reported and DIF-adjusted health profiles for each of the EQ-5D-5L dimensions). The fact that there is less variation in DIF-adjusted profiles is a limitation of the approach used, since dispersion of simulated responses will be less than that of the self-reported data by construction (Jones et al., 2011). We return to this point in the discussion. Nevertheless, we are still able to observe significant variations across unadjusted and DIF-adjusted indices.

[7] Since both models are estimated using the same location and scale restrictions (i.e. the coefficient on the constant is restricted to zero and variance of mean equation constrained to 1). Positive (negative) coefficients are increasing (decreasing) in health.

Figure 3 – Original and DIF-adjusted EQ-5D scores across sub-groups (with bootstrapped 95% confidence intervals)

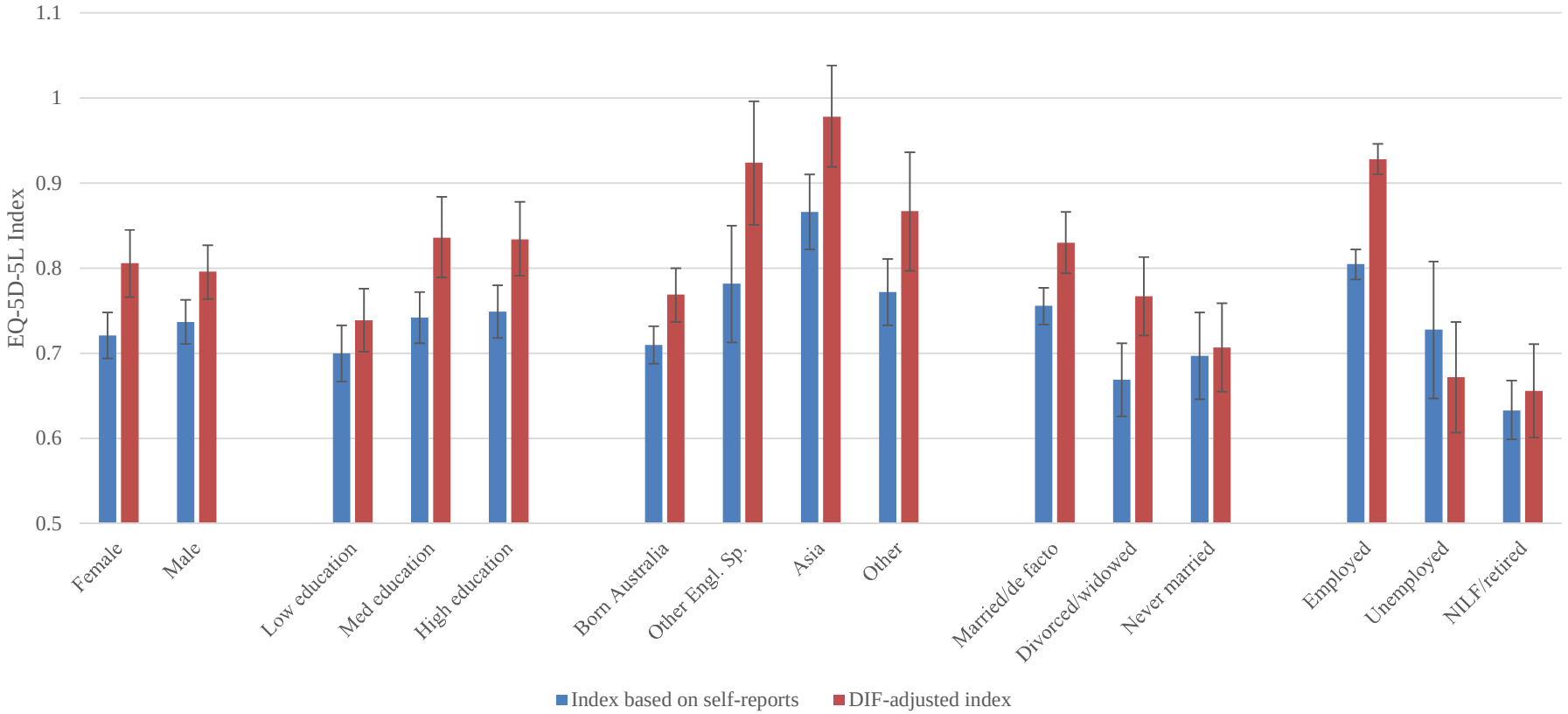


Table 5 – Group differences in unadjusted and DIF-adjusted EQ-5D-5L indices

	Reported			DIF-adjusted		
	Difference	*p*-value[a]	Above MID[b]?	Difference	*p*-value[a]	Above MID[b]?
Gender	0.016	0.403	No	-0.010	0.700	No
Education						
High and low	0.049	0.031	No	0.095	0.001	Yes
Medium and low	0.042	0.067	No	0.097	0.001	Yes
High and medium	0.008	0.727	No	-0.002	0.943	No
Country of birth						
Australia & other English speaking	0.072	0.051	No	0.155	<0.001	Yes
Australia & Asia	0.156	<0.001	Yes	0.210	<0.001	Yes
Australia & other	0.062	0.006	No	0.098	0.014	Yes
Asia & other English speaking	0.084	0.045	Yes	0.055	0.266	No
Asia & other	0.094	0.002	Yes	0.111	0.023	Yes
Other English speaking & other	0.010	0.815	No	0.057	0.286	No
Marital status						
Married & divorced	0.086	<0.001	Yes	0.063	0.034	No
Married & single	0.059	0.037	No	0.123	<0.001	Yes
Divorced & single	-0.028	0.406	No	0.060	0.100	No
Employment status						
Employed & unemployed	0.077	0.067	Yes	0.256	<0.001	Yes
Employed & NILF	0.171	<0.001	Yes	0.273	<0.001	Yes
Unemployed & NILF	0.094	0.036	Yes	0.017	0.539	No

[a]Calculated from boot-strapped standard errors, 1,000 replications.
[b]Based on a minimally important difference (MID) of 0.074 (Walters and Brazier, 2005).
Note: Based on sample aged 55-65 years (n=917)

Focussing first on gender, differences across males and females are very small and insignificant both before and after adjusting for DIF, indicating that DIF had little effect on our conclusions regarding gender differences (i.e. EQ-5D scores of males minus females, are 0.016 (p-value = 0.403) and -0.010 (p-value = 0.700) pre and post adjustment, respectively). Next, looking at education, differences in EQ-5D scores between high and low education groups, increased from 0.049 (p-value = 0.031) based on self-reports to 0.095 (p-value = 0.001) based on DIF-adjustments, while for the medium and low education groups, the difference in indices increased from 0.042 (p-value = 0.067) to 0.097 (p-value = 0.943) for unadjusted and DIF-adjusted values, respectively. Notably the DIF-adjusted difference between education groups increased to a value of clinical relevance, as the differences exceed a suggested minimally important difference (MID) of 0.074 (Walters and Brazier, 2005).[8]

In relation to country of birth, the difference between people born in Australia and other English speaking countries increased from 0.072 prior to adjustment (below the MID, p-value = 0.051) to 0.155 post-adjustment (above the MID, p-value < 0.001). The difference between Australian-born respondents and those born in Asia increased substantially from 0.156 (p-value < 0.001) to 0.210 (p-value = 0.001) post adjustment, while the difference between Australian-born respondents, and those born in other non-English speaking countries increased from 0.062 (p-value = 0.006) to 0.098 (p-value = 0.014), which again is above the MID and would therefore represent a meaningful difference in a clinical setting. Differences between respondents born in Asia and English speaking countries other than Australia changed from being statistically significant at the 5% level (0.084, p-value = 0.045) to a difference that was not (0.055, p-value = 0.266).

Variations across indices for people who were married/de facto and those who were divorced/widowed decreased from 0.086 (p-value < 0.001) to 0.063 (p-value = 0.063). While the difference between people who were married and single, and the difference between respondents who were divorced or widowed and respondents who never married, increased from 0.059 (p-value = 0.037) to 0.123 (p-value < 0.001), and from -0.028 (p-value = 0.406) to 0.060 (p-value = 0.100), respectively. Notably these changes affect the rank orderings of health by marital status.

Finally, average indices also varied substantially across subgroups according to employment status, with differences increasing from 0.077 (p-value = 0.067) to 0.256 (p-value < 0.001) for the employed and unemployed, from 0.171 to 0.273 (p-value for both <

[8] We note that that the MID of Walters and Brazier (2005) was calculated using the EQ-5D-3L; unfortunately a similar indicator – based on a variety of different patient groups, is currently unavailable for the EQ-5D-5L.

0.001) for the employed and individuals NILF or retired, while decreasing from 0.094 (*p*-value = 0.036) to 0.017 (*p*-value = 0.539) for the unemployed and respondents retired or NILF.

5.4. Robustness checks

We now return to the failure of VE in the anxiety/depression dimension and examine the extent to which the violation affects our DIF-adjusted indices and biases our findings concerning group differences in EQ-5D-5L indices. To do this we follow Kapteyn et al. (2013), by replacing equation 4 of the HOPIT with equation 9 of section 4.1 (for the anxiety/depression only), where X^- contains a variable for female (i.e. the covariate which led to a failure of VE). This model arbitrarily assumes that VE holds for one vignette (i.e. the first, which does not depend on X^-), but not for the second. Thus we repeat the analysis using X^- in the equation for the first vignette as opposed to the second. We then re-estimate the DIF-adjusted EQ-5D-5L indices of Section 4.1 under both extended models.

Results are presented in Table 6 alongside those for the standard model (i.e. the model described in Section 4.1). The top half of Table 6 shows that the EQ-5D-5L indices produced by the extended models allowing for a VE violation of vignette 2 are almost identical to the results produced under the standard DIF-adjusted model (i.e. the first column), with no statistically significant differences found across any index. This is because the coefficient on the gender dummy was small and insignificant in this model (i.e. there were no significant differences across gender in the interpretation of vignette 2), and it did not affect the coefficients of the mean function. There are however some statistically significant differences across the indices produced by the model allowing for a VE violation in the first vignette and the standard model. Specifically, the indices differ for males, respondents in the lowest education group and those born in English speaking countries other than Australia, as well as respondents who have never married, and the unemployed (note however that confidence intervals across the standard and extended model always overlap). For all of these cases, the indices estimated under the model that allows for a violation of the first vignette is lower than under the standard model.

These changes affected the estimated differences across subgroups, resulting in larger differences across almost all subgroups than estimated under the standard DIF-adjusted model, such that the effects were stronger under the extended model for vignette 1. Importantly, our conclusions regarding significant differences across subgroups, and findings

in terms of MIDs do not change across subgroups when allowing for the violation of VE (with the exception of the difference between divorced/widowed and single respondents, which increased in size to a statistically significant value). We can therefore conclude that our results regarding subgroup differences in EQ-5D-5L indices are robust, particularly in terms of their qualitative interpretation and the inferences for comparing EQ-5D indices across subgroups.

Table 6 – Robustness checks – violation of vignette equivalence

	Standard model		Extended model allowing for VE violation of vignette 1		Extended model allowing for VE violation of vignette 2	
	Value	95% CI	Value	95% CI	Value	95% CI
DIF-adjusted EQ-5D-5L indices						
Female	0.806	0.766 to 0.845	0.814	0.776 to 0.852	0.806	0.766 to 0.845
Male	0.796	0.764 to 0.827	0.772***	0.738 to 0.805	0.797	0.765 to 0.829
Education						
Low	0.739	0.702 to 0.776	0.718*	0.671 to 0.765	0.739	0.703 to 0.775
Medium	0.836	0.789 to 0.884	0.837	0.796 to 0.877	0.838	0.79 to 0.886
High	0.834	0.791 to 0.878	0.833	0.794 to 0.872	0.834	0.79 to 0.879
Country of birth						
Australia	0.769	0.737 to 0.800	0.76	0.729 to 0.790	0.769	0.737 to 0.800
Other English speaking	0.924	0.851 to 0.996	0.904***	0.83 to 0.977	0.924	0.851 to 0.997
Asia	0.978	0.919 to 1.038	0.961	0.906 to 1.017	0.978	0.919 to 1.038
Other	0.867	0.797 to 0.936	0.873	0.807 to 0.939	0.871	0.802 to 0.94
Marital status						
Married	0.83	0.794 to 0.866	0.836	0.803 to 0.869	0.831	0.795 to 0.867
Divorced/separated	0.767	0.721 to 0.813	0.77	0.722 to 0.819	0.767	0.721 to 0.813
Never married	0.707	0.655 to 0.759	0.603***	0.545 to 0.661	0.707	0.655 to 0.759
Employment status						
Employed	0.928	0.910 to 0.946	0.919	0.899 to 0.939	0.928	0.911 to 0.946
Unemployed	0.672	0.607 to 0.737	0.633**	0.555 to 0.711	0.672	0.607 to 0.737
NILF	0.656	0.601 to 0.711	0.654	0.601 to 0.706	0.657	0.602 to 0.712
	Difference	95% CI	Difference	95% CI	Difference	95% CI

Group differences in indices						
Gender	-0.01	-0.059 to 0.04	-0.043	-0.095 to 0.009	-0.009	-0.057 to 0.04
Education						
High and low	0.095***	0.039 to 0.151	0.116***	0.054 to 0.178	0.095***	0.039 to 0.151
Medium and low	0.097***	0.038 to 0.157	0.119***	0.057 to 0.182	0.099***	0.04 to 0.158
High and medium	-0.002	-0.063 to 0.059	-0.003	-0.057 to 0.05	-0.004	-0.067 to 0.058
Country of Birth						
Australia & other English speaking	0.155***	0.076 to 0.234	0.144***	0.064 to 0.224	0.155***	0.076 to 0.234
Australia & Asia	0.21***	0.142 to 0.277	0.202***	0.137 to 0.267	0.21***	0.142 to 0.278
Australia & other	0.098**	0.02 to 0.176	0.113***	0.039 to 0.188	0.103***	0.025 to 0.18
Asia & other English speaking	0.055	-0.042 to 0.151	0.058	-0.037 to 0.152	0.055	-0.042 to 0.152
Asia & other	0.111**	0.015 to 0.208	0.089*	-0.002 to 0.179	0.107**	0.011 to 0.203
Other English speaking & other	0.057	-0.047 to 0.161	0.031	-0.071 to 0.133	0.052	-0.052 to 0.156
Marital status						
Married & divorced	0.063**	0.005 to 0.121	0.066**	0.007 to 0.124	0.064**	0.006 to 0.122
Married & single	0.123***	0.057 to 0.188	0.233***	0.163 to 0.304	0.124***	0.059 to 0.189
Divorced & single	0.06	-0.011 to 0.131	0.168***	0.094 to 0.241	0.06	-0.012 to 0.132
Employment status						
Employed & unemployed	0.256***	0.188 to 0.325	0.286***	0.203 to 0.369	0.256***	0.187 to 0.325
Employed & NILF	0.273***	0.213 to 0.333	0.265***	0.207 to 0.324	0.271***	0.211 to 0.331
Unemployed & NILF	0.017	-0.036 to 0.07	-0.021	-0.095 to 0.053	0.015	-0.037 to 0.068

Note: 95% confidence intervals and tests of statistically significant differences based on bootstrapped standard errors (1,000 replications). Sample aged 55-65 years (n=917).
* p<0.10, ** p<0.05, *** p<0.01.

6. Discussion

This paper reports on an exploratory analysis involving the use of anchoring vignettes to identify DIF in the EQ-5D. We demonstrate that using vignettes to appropriately identify DIF in EQ-5D reporting is possible, although at this stage only for those aged 55-65 years. This may be of use in clinical settings for health conditions or therapies that are age-dependent, or in health or household surveys that target specific groups – e.g. older individuals. We demonstrate that failure to account for DIF can lead to conclusions that are misleading when using the instrument to compare health across heterogeneous subgroups. For instance, when adjusting for DIF in a sample aged 55-65 years, we found that differences between high and low education groups, married and single individuals, and between Australian-born respondents and those born in other English speaking countries, doubled in value after adjusting for DIF, and increased in magnitude to values that would not have had relevance in a clinical settings to ones that would (based on a suggested MID). Thus, our research provides evidence that the EQ-5D should be used with caution when identifying health disparities. Similar conclusions have been drawn by other studies examining DIF in general and domain-specific self-assessments of health (e.g. Grol-Prokopczyk et al., 2011, Molina, 2016).

In the case of economic evaluations, further work is needed in order to understand whether DIF may bring into question any accepted findings, for instance, that an intervention is cost-effective. Indeed, it may well reverse or exacerbate any such (erroneous) findings (Knott et al., 2016). Such an analysis would require administering anchoring vignettes to trial participants at baseline and follow-up alongside the EQ-5D; this was beyond the scope of the current study. It may however be a worthwhile direction for future research, as it could potentially have a significant effect on the way in which decisions are informed, given that QALYs (predominantly derived using the EQ-5D, although all subjective utility instruments may be subject to DIF) underpin the basis of funding decisions made by many health technology assessment agencies throughout the world (Dolan et al., 2013).

A drawback of our study is that the EQ-5D anchoring vignettes could not be legitimately used to make group comparisons across the entire sample due to violations of VE, which occurred in all age groups other than 55 to 65 year olds. This diminished our ability to make age-related inferences regarding DIF in the EQ-5D, which could be of particular interest. For instance, we would expect that a person aged in their twenties would

attach a different meaning to what constitutes "moderate problems walking about" compared to a person aged in their sixties.

Our study does however offer several insights in terms of the identifying assumptions of the anchoring vignette approach. Using formal, rigorous tests we found that RC held across our entire sample, which contrasts against findings of other studies assessing similar dimensions, for example Bago d'Uva et al. (2011) and Kapteyn et al. (2011) - both who examined mobility. This could reflect the effort exerted in the design stage aimed at increasing RC. Following recommendations of Au and Lorgelly (2014), instructions were given before vignettes asking respondents to rate the vignettes as if it were themselves in the health states, and to imagine that the vignette persons were of a similar age as themselves. Furthermore, vignettes avoided mention of diseases which could be dependent on age. We speculate that doing so, however, may have come at the cost of VE, which did not hold for our sample at large.

Indeed, whether it is possible to satisfy both RC and VE assumptions in a wider sample remains unknown, as the two, at least to some degree, trade-off against each other. For example, we would imagine that removing the instruction to imagine the vignettes' persons being of a similar age to themselves (i.e. the respondent), and adding more information to the vignette descriptions, such as expanding on the nature of the health limitations and attaching specific characteristics to the individuals in the vignettes, would reduce the potential for ambiguity in vignette interpretations.[9] Although this in turn would likely increase the potential for a violation of RC. Further qualitative work is needed on this point. We did, however, find VE to hold for our oldest age group; therefore it may be that RC *and* VE can only realistically be achieved in samples with similar characteristics (e.g. in terms of age). That there was consistency in vignette interpretations amongst older individuals but not amongst younger individuals could perhaps be because older respondents were better able to conceptualize the unfavourable health states described, either through personal experience or the experience of their peers – thus minimising the potential for ambiguity, and therefore variation, in interpretation. This could suggest that vignettes targeting younger age groups should be designed according to health states that they, or their peers, are likely to experience, e.g. sports injuries. However we are also unable to test this hypothesis in the current study.

[9] Making the vignette genders constant across all respondents may also improve VE.

Another limitation is that of the approach used to obtain DIF-adjusted outcomes (and therefore EQ-5D indices), since dispersion of simulated responses will be less than that of the self-reported data by construction (Jones et al., 2011). Nevertheless, we were still able to observe significant differences between unadjusted and DIF-adjusted indices across subgroups. This limitation could be somewhat alleviated by including additional health variables in the mean function of the HOPIT models – these may consist, for example, of clinically measured health indicators obtained in clinical settings.[10] Recall that variables appearing in the mean functions need not necessarily appear in the threshold equations; although including health variables in the thresholds could be an interesting exercise in its own right. This may be particularly so, for example, when considering relationships between duration of illness, adaptation and reporting styles.

In summary, we have found that the use of anchoring vignettes to identify DIF in the EQ-5D is feasible, at least amongst some population groups. Our vignettes reveal that the EQ-5D is indeed subject to DIF, which is found to bias conclusions regarding inter-group comparisons. While our study has focussed specifically on the EQ-5D, DIF may also extend to other PROMs using subjective categorical scales. Given the strong reliance on PROMs in economic evaluations for HTA, the implications of DIF could be of considerable importance, not only for outcomes research, but for funding decisions in healthcare more broadly.

[10] In the current analyses, we could have potentially used our 'objective' measures collected to assess RC; though this was not feasible as they are only available for the initial survey.

References

ANGELINI, V., CAVAPOZZI, D., CORAZZINI, L. & PACCAGNELLA, O. 2014. Do Danes and Italians Rate Life Satisfaction in the Same Way? Using Vignettes to Correct for Individual‐Specific Scale Biases. *Oxford Bulletin of Economics and Statistics,* 76**,** 643-666.

AU, N. & LORGELLY, P. K. 2014. Anchoring vignettes for health comparisons: an analysis of response consistency. *Quality of Life Research***,** 1-11.

BAGO D'UVA, T., LINDEBOOM, M., O'DONNELL, O. & VAN DOORSLAER, E. 2011. Education-related inequity in healthcare with heterogeneous reporting of health. *Journal of the Royal Statistical Society Series A: Statistics in Society,* 174**,** 639-664.

BAGO D'UVA, T., O'DONNELL, O. & VAN DOORSLAER, E. 2008a. Differential health reporting by education level and its impact on the measurement of health inequalities among older Europeans. *International Journal of Epidemiology,* 37**,** 1375-1383.

BAGO D'UVA, T., VAN DOORSLAER, E., LINDEBOOM, M. & O'DONNELL, O. 2008b. Does reporting heterogeneity bias the measurement of health disparities? *Health Economics,* 17**,** 351-375.

BAGO D'UVA, T., LINDEBOOM, M., O'DONNELL, O. & VAN DOORSLAER, E. 2011. Slipping Anchor? Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity. *Journal of Human Resources,* 46**,** 875-906.

BERTONI, M. 2015. Hungry today, unhappy tomorrow? Childhood hunger and subjective wellbeing later in life. *J Health Econ,* 40**,** 40-53.

BROOKS, R. 1996. EuroQol: the current state of play. *Health policy,* 37**,** 53-72.

BURSTRÖM, K., JOHANNESSON, M. & DIDERICHSEN, F. 2001. Swedish population health-related quality of life results using the EQ-5D. *Quality of Life Research,* 10**,** 621-635.

DEVLIN, N. J. & KRABBE, P. F. 2013. The development of new research methods for the valuation of EQ-5D-5L. *The European Journal of Health Economics,* 14**,** 1.

DEVLIN, N. J., PARKIN, D. & BROWNE, J. 2010. Patient‐reported outcome measures in the NHS: new methods for analysing and reporting EQ‐5D data. *Health economics,* 19**,** 886-905.

DOLAN, P., KAVETSOS, G. & TSUCHIYA, A. 2013. Sick but satisfied: the impact of life and health satisfaction on choice between health scenarios. *Journal of health economics,* 32**,** 708-714.

EUROQOL GROUP. 2014. *How to use EQ-5D* [Online]. Available: http://www.euroqol.org/about-eq-5d/how-to-use-eq-5d.html [Accessed August 2014.

GROL-PROKOPCZYK, H., FREESE, J. & HAUSER, R. M. 2011. Using Anchoring Vignettes to Assess Group Differences in General Self-Rated Health. *Journal of Health and Social Behavior,* 52**,** 246.

GROL-PROKOPCZYK, H., VERDES-TENNANT, E., MCENIRY, M. & ISPÁNY, M. 2015. Promises and Pitfalls of Anchoring Vignettes in Health Survey Research. *Demography,* 52**,** 1703-1728.

HOPKINS, D. J. & KING, G. 2010. Improving anchoring vignettes: Designing surveys to correct interpersonal incomparability. *Public Opinion Quarterly,* 74**,** 201-222.

JONES, A. M., RICE, N., ROBONE, S. & DIAS, P. R. 2011. Inequality and polarisation in health systems' responsiveness: a cross-country analysis. *Journal of health economics,* 30**,** 616-625.

JÜRGES, H. & WINTER, J. 2011. Are anchoring vignettes ratings sensitive to vignette age and sex? *Health Economics***,** n/a-n/a.

JÜRGES, H. & WINTER, J. 2013. Are anchoring vignettes ratings sensitive to vignette age and sex? *Health economics,* 22**,** 1-13.

KAPTEYN, A., SMITH, J., VAN SOEST, A. & VONKOVA, H. 2011. Anchoring vignettes and response consistency. *RAND Working Paper Series WR-840*.

KAPTEYN, A., SMITH, J. P. & VAN SOEST, A. 2007. Vignettes and self-reports of work disability in the United States and the Netherlands. *The American Economic Review,* 97**,** 461-473.

KAPTEYN, A., SMITH, J. P. & VAN SOEST, A. 2013. Are Americans really less happy with their incomes? *Review of Income and Wealth,* 59**,** 44-65.

KIND, P., DOLAN, P., GUDEX, C. & WILLIAMS, A. 1998. Variations in population health status: results from a United Kingdom national questionnaire survey. *BMJ,* 316**,** 736-741.
KING, G., MURRAY, C. J. L., SALOMON, J. A. & TANDON, A. 2004. Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review,* 98**,** 191-207.
KNOTT, R. J., BLACK, N., HOLLINGSWORTH, B. & LORGELLY, P. K. 2016. Response Scale Heterogeneity in the EQ5D. *Health economics*.
KO, Y. & COONS, S. J. 2006. Self-reported chronic conditions and EQ‐5D index scores in the US adult population. *Current Medical Research and Opinion®,* 22**,** 2065-2071.
KRISTENSEN, N. & JOHANSSON, E. 2008. New evidence on cross-country differences in job satisfaction using anchoring vignettes. *Labour Economics,* 15**,** 96-117.
LUBETKIN, E., JIA, H., FRANKS, P. & GOLD, M. 2005. Relationship Among Sociodemographic Factors, Clinical Conditions, and Health-related Quality of Life: Examining the EQ-5D in the U.S. General Population. *Quality of Life Research,* 14**,** 2187-2196.
LUO, N., JOHNSON, J. A., SHAW, J. W., FEENY, D. & COONS, S. J. 2005. Self-reported health status of the general adult US population as assessed by the EQ-5D and Health Utilities Index. *Medical Care,* 43**,** 1078-1086.
MAHESWARAN, H., PETROU, S., REES, K. & STRANGES, S. 2012. Estimating EQ-5D utility values for major health behavioural risk factors in England. *Journal of Epidemiology and Community Health*.
MOLINA, T. 2016. Reporting Heterogeneity and Health Disparities Across Gender and Education Levels: Evidence From Four Countries. *Demography***,** 1-29.
MURRAY, C. J. L., TANDON, A., SALOMON, J. A., MATHERS, C. D. & SADANA, R. 2002. New approaches to enhance cross-population comparability of survey results. *Summary measures of population health: concepts, ethics, measurement and applications.* Geneva: World Health Organization.
NORMAN, R., CRONIN, P. & VINEY, R. 2013. A pilot discrete choice experiment to explore preferences for EQ-5D-5L health states. *Applied health economics and health policy,* 11**,** 287-298.
PERACCHI, F. & ROSSETTI, C. 2013. The heterogeneous thresholds ordered response model: Identification and inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 176**,** 703-722.
PROSSER, L. A., STINNETT, A. A., GOLDMAN, P. A., WILLIAMS, L. W., HUNINK, M. G., GOLDMAN, L. & WEINSTEIN, M. C. 2000. Cost-effectiveness of cholesterol-lowering therapies according to selected patient characteristics. *Annals of Internal Medicine,* 132**,** 769-779.
SALOMON, J. A., PATEL, A., NEAL, B., GLASZIOU, P., GROBBEE, D. E., CHALMERS, J. & CLARKE, P. M. 2011. Comparability of Patient-reported Health Status: Multicountry Analysis of EQ-5D Responses in Patients With Type 2 Diabetes. *Medical Care,* 49**,** 962.
SØLTOFT, F., HAMMER, M. & KRAGH, N. 2009. The association of body mass index and health-related quality of life in the general population: data from the 2003 Health Survey of England. *Quality of Life Research,* 18**,** 1293-1299.
SULLIVAN, P. W. & GHUSHCHYAN, V. 2006. Preference-based EQ-5D index scores for chronic conditions in the United States. *Medical Decision Making,* 26**,** 410-420.
SUN, S., CHEN, J., JOHANNESSON, M., KIND, P., XU, L., ZHANG, Y. & BURSTRÖM, K. 2011. Population health status in China: EQ-5D results, by age, sex and socio-economic status, from the National Health Services Survey 2008. *Quality of Life Research,* 20**,** 309-320.
WALTERS, S. J. & BRAZIER, J. E. 2005. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. *Quality of Life Research,* 14**,** 1523-1532.
WHYNES, D. K., SPRIGG, N., SELBY, J., BERGE, E., BATH, P. M. & INVESTIGATORS, E. 2013. Testing for Differential Item Functioning within the EQ-5D. *Medical Decision Making,* 33**,** 252-260.

Appendix A – Gender-specific vignettes (i.e. vignettes were presented according to the gender of the respondent)

Vignette 1:
REBECCA/ROB is able to walk distances of up to 500 metres without any problems but feels puffed and tired after walking one kilometre or walking up more than one flight of stairs. She/he is able to wash, dress and groom her/himself, but it requires some effort due to an injury from an accident one year ago. Her/his injury causes her/him to stay home from work or social activities about once a month. Rebecca/Rob feels some stiffness and pain in her/his right shoulder most days however her/his symptoms are usually relieved with low doses of medication, stretching and massage. She/he feels happy and enjoys things like hobbies or social activities around half of the time. The rest of the time she/he worries about the future and feels depressed a couple of days a month.

Vignette 2:
CHRISTINE/CHRIS is suffering from an injury which causes her/him a considerable amount of pain. She/he can walk up to a distance of 50 metres without any assistance, but struggles to walk up and down stairs. She/he can wash her/his face and comb her/his hair, but has difficulty washing her/his whole body without help. She/he needs assistance with putting clothes on the lower half of her/his body. Since having the injury Christine/Chris can no longer cook or clean the house her/himself, and needs someone to do the grocery shopping for her/him. The injury has caused her/him to experience back pain every day and she/he is unable to stand or sit for more than half an hour at a time. She/he is depressed nearly every day and feels hopeless. She/he also has a low self-esteem and feels that she/he has become a burden.

Appendix B: Additional tables

Table B1 – Sample characteristics

	Entire sample (N = 4095)		Sample used in DIF analysis (N = 914)	
	Mean	St. Dev.	Mean	Std.Dev.
Female	0.518	0.500	0.497	0.500
Age	42.2	12.9	59.8	3.3
Highest level of education:				
University (high)	0.409	0.492	0.309	0.462
Certificate or diploma (medium)	0.298	0.458	0.330	0.471
Year 12 or less (low)	0.293	0.455	0.361	0.481
Country of birth				
Australia	0.750	0.433	0.756	0.430
Other English speaking	0.051	0.221	0.069	0.253
Asia	0.091	0.287	0.036	0.187
Other	0.107	0.310	0.139	0.346
Employment status				
Employed	0.682	0.466	0.528	0.499
Unemployed	0.103	0.304	0.053	0.223
Not in labour force/retired	0.215	0.411	0.419	0.494
EQ-5D-5L Index	0.750	0.270	0.729	0.286
EQ-5D-5L dimensions – average rating				
Mobility	1.377	0.744	1.579	0.884
Self-care	1.162	0.526	1.194	0.564
Usual activities	1.400	0.774	1.537	0.842
Pain	1.889	1.027	2.103	0.989
Anxiety/depression	1.864	0.943	1.736	0.971
Vignette 1 - average rating				
Mobility	2.407	0.745	2.515	0.724
Self-care	2.136	0.757	2.144	0.760
Usual activities	2.366	0.731	2.403	0.707
Pain	2.579	0.686	2.606	0.633
Anxiety/depression	2.435	0.812	2.470	0.773
Vignette 2 - average rating				
Mobility	3.590	0.761	3.700	0.615
Self-care	3.645	0.831	3.777	0.707
Usual activities	3.962	0.889	4.100	0.752
Pain	3.896	0.808	4.005	0.643
Anxiety/depression	4.005	0.921	4.081	0.768

Vignette order violations				
Mobility	0.034	0.182	0.021	0.143
Self-care	0.031	0.174	0.016	0.127
Usual activities	0.029	0.168	0.027	0.163
Pain	0.030	0.171	0.014	0.118
Anxiety/depression	0.034	0.182	0.021	0.143

Table B2 – Estimated coefficients of ordered probit (OP) and HOPIT models

	Mobility		Personal care		Usual activities		Pain/ Discomfort		Anxiety/ Depression	
	OP	HOPIT	OP	HOPIT	OP	HOPIT	OP	HOPIT	OP	HOPIT
Female	0.123	0.133	0.263**	0.425***	0.085	0.143	-0.001	0.04	-0.097	0.08
	(0.082)	(0.088)	(0.110)	(0.109)	(0.083)	(0.089)	(0.073)	(0.082)	(0.077)	(0.084)
Education (reference low)										
Medium	0.141	0.226**	0.282**	0.469***	0.034	0.125	0.094	0.022	0.032	0.151
	(0.097)	(0.106)	(0.132)	(0.133)	(0.098)	(0.106)	(0.087)	(0.098)	(0.093)	(0.102)
High	0.082	0.132	0.076	0.299**	0.078	0.184*	0.072	0.038	-0.025	0.13
	(0.100)	(0.107)	(0.130)	(0.129)	(0.102)	(0.109)	(0.090)	(0.101)	(0.096)	(0.103)
Country of Birth (reference Australia)										
Other Engl. sp. country	-0.044	-0.054	0.195	0.27	0.092	0.273	0.400***	0.334*	0.307*	0.489**
	(0.161)	(0.183)	(0.233)	(0.245)	(0.166)	(0.200)	(0.149)	(0.184)	(0.160)	(0.194)
Asia	0.934***	0.845**	4.23	5.721	0.725**	0.727**	0.651***	0.507*	0.115	0.237
	(0.316)	(0.346)	(144.903)	(107.623)	(0.293)	(0.333)	(0.211)	(0.263)	(0.215)	(0.262)
Other country	0.258**	0.237*	0.319*	0.397**	0.269**	0.297**	0.246**	0.336***	0.068	0.052
	(0.120)	(0.135)	(0.172)	(0.180)	(0.122)	(0.139)	(0.105)	(0.128)	(0.111)	(0.126)
Marital status (reference never married)										
Married/de facto	0.164	0.406***	0.099	0.634***	0.197	0.482***	0.190*	0.186**	0.178	0.415***
	(0.124)	(0.094)	(0.167)	(0.111)	(0.127)	(0.096)	(0.112)	(0.090)	(0.118)	(0.092)
Divorced/Widowed	0.028	0.224*	-0.184	0.278*	-0.079	0.165	-0.066	-0.027	-0.144	0.096
	(0.141)	(0.124)	(0.185)	(0.146)	(0.142)	(0.126)	(0.127)	(0.118)	(0.134)	(0.120)
Employment status (reference NILF/retired)										
Employed	0.679***	0.950***	0.708***	0.963***	0.696***	0.891***	0.514***	0.683***	0.363***	0.596***
	(0.085)	(0.095)	(0.115)	(0.118)	(0.086)	(0.096)	(0.076)	(0.087)	(0.080)	(0.088)
Unemployed	0.136	0.592***	0.407	0.827***	0.336*	0.793***	0.237	0.478**	0.398**	0.556***
	(0.178)	(0.205)	(0.254)	(0.265)	(0.183)	(0.230)	(0.167)	(0.199)	(0.180)	(0.203)

Vignette 2 constant	-0.994***	-1.100***	-1.311***	-1.246***	-1.219***
	(0.059)	(0.100)	(0.075)	(0.058)	(0.063)
σ_v	0.513***	0.469***	0.537***	0.524***	0.557***
	(0.028)	(0.042)	(0.030)	(0.023)	(0.027)
OP threshold constants					
τ_1	-2.150***	-2.354***	-2.126***	-1.698***	-1.841***
	(0.215)	(0.286)	(0.207)	(0.154)	(0.157)
τ_2	-1.102***	-1.668***	-1.297***	-0.844***	-1.318***
	(0.146)	(0.201)	(0.151)	(0.130)	(0.140)
τ_3	-0.419***	-1.061***	-0.526***	-0.036	-0.618***
	(0.140)	(0.183)	(0.142)	(0.128)	(0.135)
τ_4	0.322**	-0.476***	0.249*	1.062***	0.201
	(0.139)	(0.178)	(0.141)	(0.131)	(0.134)

* p<0.10, ** p<0.05, *** p<0.01.
Note: Standard errors in parentheses. Sample aged 55-65 years (n=917).

Table B3 – HOPIT estimates for all thresholds

	Mobility	Self care	Usual activities	Pain/ Discomfort	Anxiety/ Depression	Mobility	Self care	Usual activities	Pain/ Discomfort	Anxiety/ Depression
	First threshold (between "extreme problems" and "severe problems")					*Second threshold (between "severe problems" and "moderate problems")*				
Female	-0.165*	-0.005	0.059	0.131***	0.035	0.070	0.040	-0.044	-0.060	-0.011
	(0.087)	(0.052)	(0.046)	(0.050)	(0.047)	(0.073)	(0.070)	(0.069)	(0.060)	(0.070)
Education (base category low)										
Medium	-0.128	-0.088	0.014	-0.109*	0.047	0.069	-0.036	-0.149*	0.007	-0.092
	(0.095)	(0.061)	(0.054)	(0.057)	(0.055)	(0.080)	(0.082)	(0.081)	(0.069)	(0.081)
High	-0.251**	-0.168**	-0.073	-0.142**	-0.03	0.104	0.074	-0.097	-0.018	-0.077
	(0.107)	(0.067)	(0.057)	(0.061)	(0.058)	(0.089)	(0.085)	(0.085)	(0.074)	(0.085)
Country of Birth (ref. Australia)										
Oth English speaking	0.099	0.125	-0.097	0.188**	0.119	-0.059	-0.194	0.118	-0.403***	-0.146
	(0.160)	(0.095)	(0.094)	(0.089)	(0.088)	(0.137)	(0.145)	(0.129)	(0.142)	(0.140)
Asia	0.168	0.037	0.025	0.055	0.02	-0.095	-0.039	0.062	0.005	-0.13
	(0.105)	(0.073)	(0.065)	(0.070)	(0.066)	(0.095)	(0.098)	(0.097)	(0.083)	(0.101)
Other	0.399**	0.159	0.142	0.201	0.118	-0.282	-0.221	-0.185	-0.351*	-0.182
	(0.179)	(0.133)	(0.121)	(0.126)	(0.123)	(0.182)	(0.196)	(0.212)	(0.201)	(0.195)
Marital status (ref. never married)										
Married/de facto	-0.335***	-0.165**	-0.005	-0.063	0.008	0.233**	0.193*	0.069	-0.044	0.143
	(0.103)	(0.074)	(0.070)	(0.074)	(0.073)	(0.100)	(0.106)	(0.107)	(0.089)	(0.113)
Divorced/widowed	-0.259**	-0.123	0.066	-0.034	0.092	0.238**	0.163	0.040	0.000	0.182
	(0.123)	(0.084)	(0.079)	(0.084)	(0.081)	(0.113)	(0.121)	(0.122)	(0.099)	(0.121)
Employment status (ref. NILF)										
Employed	-0.009	-0.032	-0.074	-0.044	-0.087*	0.001	0.036	0.108	0.074	0.128*
	(0.084)	(0.053)	(0.048)	(0.051)	(0.048)	(0.072)	(0.070)	(0.072)	(0.061)	(0.072)

Unemployed	-0.333	-0.127	0.018	-0.023	-0.269**	0.308*	0.106	-0.231	-0.003	0.345**
	(0.265)	(0.128)	(0.102)	(0.113)	(0.120)	(0.170)	(0.165)	(0.190)	(0.141)	(0.144)
Constant	-1.517***	-1.452***	-1.578***	-1.649***	-1.518***	-0.078	-0.393***	-0.294**	-0.004	-0.401***
	(0.136)	(0.148)	(0.113)	(0.105)	(0.107)	(0.118)	(0.148)	(0.129)	(0.107)	(0.133)
	Third threshold (between "moderate problems" and "slight problems")					*Fourth threshold (between "slight problems" and "no problems")*				
Female	-0.010	-0.062	-0.110	-0.143**	-0.08	0.079	0.049	0.086	0.119*	0.243***
	(0.060)	(0.072)	(0.068)	(0.058)	(0.068)	(0.076)	(0.073)	(0.068)	(0.070)	(0.070)
Education (base category low)										
Medium	0.013	0.168*	0.049	0.085	0.014	0.118	0.111	0.175**	-0.057	0.123
	(0.074)	(0.086)	(0.083)	(0.070)	(0.083)	(0.089)	(0.086)	(0.081)	(0.085)	(0.084)
High	0.171**	0.202**	0.170**	0.109	0.099	-0.025	0.080	0.122	0.080	0.178**
	(0.073)	(0.089)	(0.083)	(0.072)	(0.082)	(0.095)	(0.090)	(0.086)	(0.085)	(0.083)
Country of Birth (ref. Australia)										
Oth English speaking	0.007	-0.058	0.121	0.117	-0.021	-0.075	0.2	0.144	-0.065	0.274**
	(0.114)	(0.145)	(0.123)	(0.108)	(0.139)	(0.158)	(0.132)	(0.149)	(0.146)	(0.138)
Asia	-0.206**	-0.04	-0.132	0.000	-0.066	0.065	0.077	0.037	0.062	0.148
	(0.096)	(0.104)	(0.107)	(0.084)	(0.100)	(0.102)	(0.102)	(0.099)	(0.102)	(0.093)
Other	0.049	0.036	0.079	-0.032	-0.032	-0.383	0.167	-0.091	-0.067	0.242
	(0.151)	(0.177)	(0.172)	(0.164)	(0.182)	(0.281)	(0.199)	(0.215)	(0.174)	(0.178)
Marital status (ref. never married)										
Married/de facto	0.078	-0.021	-0.014	0.063	-0.161*	0.112	0.118	0.083	0.045	-0.058
	(0.091)	(0.105)	(0.098)	(0.092)	(0.096)	(0.119)	(0.114)	(0.104)	(0.107)	(0.104)
Divorced/widowed	-0.066	0.000	-0.119	0.037	-0.282**	0.066	-0.036	0.071	0.033	-0.086

	(0.106)	(0.119)	(0.115)	(0.102)	(0.112)	(0.133)	(0.130)	(0.118)	(0.122)	(0.119)
Employment status (ref. NILF)										
Employed	0.158**	0.178**	0.073	0.113*	0.076	0.242***	0.063	0.164**	0.141*	0.227***
	(0.063)	(0.076)	(0.072)	(0.061)	(0.070)	(0.077)	(0.074)	(0.071)	(0.073)	(0.071)
Unemployed	0.248*	0.366**	0.269*	0.154	-0.068	0.251	0.077	0.434***	0.322**	0.195
	(0.130)	(0.143)	(0.141)	(0.128)	(0.162)	(0.186)	(0.182)	(0.162)	(0.163)	(0.154)
Constant	-0.455***	-0.677***	-0.351***	-0.306***	-0.229**	-0.617***	-0.693***	-0.494***	-0.119	-0.567***
	(0.111)	(0.148)	(0.120)	(0.106)	(0.116)	(0.132)	(0.143)	(0.120)	(0.115)	(0.122)

* p<0.10, ** p<0.05, *** p<0.01.
Note: Standard errors in parentheses. Sample aged 55-65 years (n=917)

Table B4 - Reported and DIF-adjusted EQ-5D-5L profiles – proportion of respondents in each category

		Gender		Age group				Education			Country of birth				Employment status			Marital status		
	Limitations	M	F	20-34	35-44	45-54	55+	Low	Med	High	Aus	Oth Engl.	Asia	Other	Empl.	Unemp.	NILF	Marr/d.f.	Div/wid	Never marr.
Mobility																				
Reported	No	0.732	0.763	0.834	0.781	0.701	0.630	0.705	0.722	0.797	0.726	0.771	0.898	0.766	0.804	0.725	0.582	0.768	0.624	0.755
	Slight	0.159	0.156	0.106	0.155	0.177	0.216	0.175	0.178	0.129	0.171	0.110	0.075	0.155	0.135	0.180	0.215	0.153	0.182	0.155
	Mod	0.082	0.057	0.051	0.047	0.085	0.104	0.079	0.074	0.059	0.075	0.071	0.027	0.059	0.048	0.083	0.129	0.058	0.125	0.071
	Sever	0.022	0.022	0.006	0.015	0.028	0.046	0.037	0.023	0.010	0.025	0.029	0	0.016	0.010	0.012	0.065	0.017	0.068	0.012
	Extr.	0.006	0.003	0.003	0.001	0.009	0.004	0.004	0.003	0.005	0.004	0.019	0	0.005	0.003	0	0.009	0.004	0.002	0.006
DIF-adj.	No	0.870	0.954	0.987	0.955	0.910	0.765	0.842	0.930	0.953	0.896	0.962	1.000	0.939	1.000	0.993	0.602	0.949	0.757	0.901
	Slight	0.130	0.046	0.013	0.045	0.090	0.235	0.158	0.070	0.047	0.104	0.038	0	0.061	0	0.007	0.398	0.051	0.243	0.099
	Mod	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Sever	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Extr.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Self-care																				
Reported	No	0.873	0.912	0.913	0.906	0.874	0.870	0.863	0.903	0.907	0.884	0.900	0.954	0.905	0.918	0.912	0.807	0.905	0.812	0.902
	Slight	0.073	0.057	0.046	0.060	0.080	0.082	0.078	0.061	0.058	0.071	0.062	0.024	0.059	0.049	0.064	0.113	0.057	0.112	0.062
	Mod	0.036	0.026	0.028	0.028	0.036	0.035	0.045	0.025	0.025	0.034	0.033	0.016	0.020	0.024	0.021	0.057	0.028	0.061	0.026
	Sever	0.014	0.004	0.010	0.004	0.009	0.011	0.010	0.008	0.008	0.010	0.005	0.005	0.007	0.008	0.002	0.015	0.009	0.015	0.005
	Extr.	0.004	0.001	0.003	0.002	0.001	0.002	0.004	0.002	0.001	0.002	0	0	0.009	0.001	0	0.008	0.001	0	0.005
DIF-adj.	No	0.964	1.000	0.990	0.989	0.984	0.963	0.947	1.000	0.995	0.978	0.995	1.000	0.995	1.000	1.000	0.918	1.000	0.956	0.956
	Slight	0.036	0	0.010	0.011	0.016	0.037	0.053	0	0.005	0.022	0.005	0	0.005	0	0	0.082	0	0.044	0.044
	Mod	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Sever	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Extr.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Usual activities																				

Reported	No	0.737	0.739	0.809	0.785	0.674	0.644	0.701	0.708	0.787	0.719	0.714	0.892	0.750	0.802	0.701	0.553	0.769	0.571	0.741
	Slight	0.158	0.161	0.122	0.129	0.192	0.217	0.173	0.174	0.139	0.168	0.176	0.086	0.155	0.136	0.201	0.214	0.147	0.233	0.156
	Mod	0.072	0.074	0.052	0.058	0.093	0.102	0.088	0.084	0.055	0.080	0.076	0.019	0.070	0.048	0.076	0.151	0.061	0.142	0.072
	Sever	0.026	0.020	0.013	0.020	0.033	0.032	0.029	0.028	0.016	0.026	0.024	0.003	0.018	0.011	0.017	0.066	0.019	0.042	0.025
	Extr.	0.007	0.006	0.004	0.008	0.008	0.005	0.008	0.007	0.004	0.007	0.010	0	0.007	0.003	0.005	0.016	0.005	0.013	0.006
DIF-adj.	No	0.861	0.925	0.968	0.940	0.871	0.758	0.815	0.897	0.949	0.874	0.943	1	0.918	1.000	0.976	0.519	0.946	0.689	0.866
	Slight	0.139	0.075	0.032	0.060	0.129	0.242	0.185	0.103	0.051	0.126	0.057	0	0.082	0	0.024	0.481	0.054	0.311	0.134
	Mod	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Sever	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Extr.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pain																				
Reported	No	0.451	0.399	0.534	0.458	0.339	0.309	0.403	0.384	0.469	0.403	0.467	0.535	0.457	0.468	0.396	0.298	0.434	0.288	0.459
	Slight	0.358	0.383	0.330	0.369	0.409	0.396	0.348	0.378	0.382	0.367	0.343	0.414	0.377	0.382	0.367	0.338	0.384	0.355	0.350
	Mod	0.117	0.153	0.100	0.115	0.149	0.197	0.160	0.156	0.103	0.153	0.114	0.040	0.102	0.110	0.178	0.196	0.129	0.197	0.124
	Sever	0.062	0.049	0.032	0.041	0.079	0.081	0.074	0.067	0.033	0.062	0.038	0.011	0.057	0.032	0.045	0.135	0.041	0.127	0.056
	Extr.	0.012	0.016	0.004	0.017	0.024	0.018	0.015	0.016	0.013	0.015	0.038	0	0.007	0.008	0.014	0.033	0.011	0.034	0.012
DIF-adj.	No	0.220	0.399	0.497	0.358	0.176	0.125	0.116	0.197	0.538	0.202	0.590	0.828	0.518	0.428	0.201	0	0.424	0.008	0.198
	Slight	0.749	0.595	0.503	0.641	0.782	0.836	0.842	0.786	0.461	0.774	0.410	0.172	0.482	0.572	0.799	0.916	0.576	0.888	0.781
	Mod	0.031	0.006	0	0.001	0.042	0.039	0.043	0.016	0.002	0.024	0	0	0	0	0	0.084	0	0.104	0.022
	Sever	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Extr.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Anxiety/Depression																				
Reported	No	0.496	0.420	0.429	0.439	0.437	0.535	0.444	0.443	0.476	0.429	0.524	0.594	0.500	0.493	0.353	0.391	0.504	0.395	0.381
	Slight	0.285	0.319	0.296	0.338	0.302	0.276	0.272	0.310	0.320	0.307	0.271	0.282	0.307	0.321	0.256	0.269	0.304	0.258	0.319
	Mod	0.144	0.179	0.185	0.153	0.175	0.128	0.190	0.153	0.149	0.179	0.143	0.091	0.116	0.135	0.268	0.200	0.144	0.216	0.180
	Sever	0.051	0.052	0.054	0.045	0.064	0.040	0.066	0.057	0.036	0.056	0.038	0.013	0.055	0.036	0.069	0.092	0.035	0.082	0.074
	Extr.	0.024	0.030	0.036	0.025	0.022	0.021	0.028	0.037	0.019	0.029	0.024	0.019	0.023	0.016	0.055	0.049	0.014	0.049	0.046

DIF-adj.	No	0.412	0.422	0.236	0.511	0.501	0.505	0.221	0.431	0.548	0.339	0.757	0.745	0.523	0.588	0.107	0.027	0.620	0.195	0.075
	Slight	0.535	0.546	0.698	0.437	0.465	0.495	0.680	0.537	0.444	0.607	0.243	0.255	0.468	0.412	0.791	0.829	0.380	0.761	0.795
	Mod	0.052	0.032	0.066	0.053	0.034	0	0.099	0.032	0.007	0.054	0	0	0.009	0	0.102	0.144	0	0.044	0.129
	Sever	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Extr.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Note: Sample aged 55-65 years (n=917)