Accepted Manuscript

Differential item functioning in quality of life measurement: An analysis using anchoring vignettes

Rachel J. Knott, Paula K. Lorgelly, Nicole Black, Bruce Hollingsworth

PII: S0277-9536(17)30514-2

DOI: 10.1016/j.socscimed.2017.08.033

Reference: SSM 11373

To appear in: *Social Science & Medicine*

Received Date: 12 January 2017

Revised Date: 14 August 2017

Accepted Date: 24 August 2017

Please cite this article as: Knott, R.J., Lorgelly, P.K., Black, N., Hollingsworth, B., Differential item functioning in quality of life measurement: An analysis using anchoring vignettes, *Social Science & Medicine* (2017), doi: 10.1016/j.socscimed.2017.08.033.

Differential item functioning in quality of life measurement: An analysis using anchoring vignettes

Manuscript Number: SSM-D-17-00108

Rachel J Knott[a]*, Paula K Lorgelly[a,b], Nicole Black[a] and Bruce Hollingsworth[c]

[a] Centre for Health Economics, Monash University, Australia
[b] Office of Health Economics, UK
[c] Division of Health Research, Lancaster University, UK

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.

**Corresponding author: rachel.knott@monash.edu*
15 Innovation Walk, Clayton, Victoria, Australia, 3800.
Ph: +61 3 9905 0772; Fax: +61 3 9905 834

Differential item functioning in quality of life measurement: An analysis using anchoring vignettes

Abstract

Systematic differences in the ways that people use and interpret response categories (differential item functioning, *DIF*) can introduce bias when using self-assessments to compare health or quality of life across heterogeneous groups. This paper reports on an exploratory analysis involving the use of anchoring vignettes to identify differential item functioning (DIF) in a commonly used measure for assessing health-related quality of life - namely the EQ-5D. Using data from a bespoke (i.e. custom) survey that recruited a representative sample of 4,300 respondents from the general Australian population in 2014 and 2015, we find that the assumptions of response consistency (RC) and vignette equivalence (VE) hold in a sub-sample of respondents aged 55 to 65 years (n=914), which demonstrates that vignettes can appropriately identify DIF in EQ-5D reporting for this age group. 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. We also provide several important insights in terms of the identifying assumptions of RC and VE. We conclude that the implications of DIF could be of considerable importance, not only for outcomes research, but for funding decisions in healthcare more broadly given the strong reliance on patient-reported outcome measures in economic evaluations for health technology assessment.

Word count: 7,946
Figure count: 3
Table count: 5

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

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 commonly 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, Murray, Salomon, & Tandon, 2004). DIF has been shown to exist across a range of subject areas, including in other self-reported measures of health (Bago d'Uva, van Doorslaer, Lindeboom, & O'Donnell, 2008; Dowd & Todd, 2011; Grol-Prokopczyk, Freese, & Hauser, 2011; Hirve et al., 2013; Peracchi & Rossetti, 2012) and wellbeing , but has largely been overlooked in the case of preference-based Health-Related Quality of Life (HRQoL) measures, and in particular, the increasingly popular Patient Reported Outcome Measures (PROMs).

The most commonly used PROM is the EuroQol's EQ-5D, which asks respondents to describe their health on five different dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression (Brooks, 1996). 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).

In both population health and economic evaluation settings the instrument is often used to compare 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 (Sullivan & Ghushchyan, 2006), behavioural risk factors (Devlin, Parkin, & Browne, 2010; Søltoft, Hammer, & Kragh, 2009), and socio-demographic characteristics such as age, gender and socio-economic status (Burström, Johannesson, & Diderichsen, 2001; Lubetkin, Jia, Franks, & Gold, 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 (or other measures of HRQoL) 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). 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, Tandon, Salomon, Mathers, & Sadana, 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,life, and health satisfaction (Angelini, Cavapozzi, Corazzini, & Paccagnella, 2014; Bertoni, 2015; Crane, Rissel, Greaves, & Gebel, 2016; Kapteyn, Smith, & Van Soest, 2013; Kristensen & Johansson, 2008) and general and specific dimensions of health (Bago d'Uva, Lindeboom, O'Donnell, & van Doorslaer, 2011; Bago d'Uva, van Doorslaer, et al., 2008; Dowd & Todd, 2011; Grol-Prokopczyk et al., 2011; Hirve et al., 2013; Molina, 2016; Peracchi & Rossetti, 2012).

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, Verdes-Tennant, McEniry, & Ispány, 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 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 (Grol-Prokopczyk et al., 2011; Kristensen & Johansson, 2008); however 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, Lindeboom, O'Donnell, & van Doorslaer, 2011; Grol-Prokopczyk et al., 2015; Peracchi & Rossetti, 2013).

This study investigates (1) whether the anchoring vignette approach can appropriately be used to identify DIF in the EQ-5D-5L, and (2) the presence and impact of DIF on inter-group comparisons using EQ-5D indices. 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 pilot study. Using a larger sample of respondents from the general Australian population, we 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). 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 describes our data and vignettes design in detail, while Section 4 outlines the methodology we use to test the identifying assumptions, and to test and adjust for DIF in the EQ-5D-5L. Section 5 presents our results and robustness checks and 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.

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. 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, i.e. Group

B's ratings are more favourable than Group A's ratings for both vignettes. Vignettes can therefore help to identify differential reporting behaviour.

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 data from two online surveys of the general Australian population. 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 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, which were included to proxy 'objective' health in the tests for RC proposed by Bago d'Uva et al. (2011). For the mobility dimension we adapted the detailed 'objective' set of mobility questions used 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 Monash University Human Research Ethics Committee (reference number CF12/0828-2012000367).

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. 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 [INSERT LINK TO ONLINE FILE] of the online appendix). 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 & 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 & King, 2010)), and names were selected from a list of the most popular names since the 1940s that appeared across several decades (Jürges & Winter, 2013). 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, O'Donnell, & van Doorslaer, 2008; Bago d'Uva, van Doorslaer, et al., 2008; Grol-Prokopczyk et al., 2011; Molina, 2016). Age is considered in terms of four categories: 20- 34; 35- 44; 45- 54 and 55- 65 at the time of survey. 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/ de facto; 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 E1 of online Appendix E [INSERT LINK TO ONLINE FILE].

4. Methodological approach

For a detailed description of our econometric approach, refer to Appendix B [INSERT LINK TO ONLINE FILE]. In brief, we adopted the hierarchical ordered probit (HOPIT) model with anchoring vignettes approach introduced by King et al. (2004). The HOPIT is an extension of the ordered probit model which allows for variation in the inter-category thresholds by modelling them as a function of covariates – the anchoring vignettes are used to identify these threshold covariates. 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.

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 conducting counterfactual simulations to obtain DIF-adjusted outcomes for each of the EQ-5D-5L dimensions. EQ-5D-5L weights (valued using an Australian discrete choice experiment (Norman, Cronin, & Viney, 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.

As mentioned above, we test the identifying assumptions of RC and VE using the methodology developed by Bago d'Uva et al. (2011). The test for RC relies on objective measures for each construct of interest (i.e. in our case, dimensions of the EQ-5D). These objective measures are used to capture variations in the self-assessments, and the covariates of the inter-category threshold equations are compared against those of the vignette ratings. RC is assessed using an LR test examining the distances between any two inter-category thresholds. The VE test involves testing for significance of individual characteristics as covariates in vignette equations, and is also conducted using LR tests. For a detailed description of both tests, see Appendix B [INSERT LINK TO ONLINE FILE], Section B.2.

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 (Bago d'Uva et al., 2011; Kapteyn et al., 2011; Peracchi & 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.

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 preliminary instructions asking respondents to imagine that the vignettes were of a similar age and background to themselves; and features of the vignette wording that made the hypothetical characters relatable to a diverse audience (i.e. gender-matching vignette characters to respondents and the avoidance of age-related health issues), may have opened up the potential for variations in the perceptions of the vignette-described health states (e.g. across age and gender), 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 could potentially 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 (i.e. we cannot conduct the test on each vignette in isolation).

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 of Appendix B [INSERT LINK TO ONLINE FILE]), which was negative and statistically significant (p=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- 65 (N =914), for which we can be reasonably certain that the vignettes are adequately identifying DIF (see Table E1 of Appendix E [INSERT LINK TO ONLINE FILE] 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- 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 for the first threshold and Table E2 of Appendix E [INSERT LINK TO ONLINE FILE] for the remaining thresholds) 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. The sources of DIF across various dimensions are discussed further in Appendix E [INSERT LINK TO ONLINE FILE].

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- 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- 65 increases from 0.729 (unadjusted) to 0.806 (DIF-adjusted) where the upward adjustment (which is visibly evident across all subgroups) reflects the movement of respondents away from the most limiting categories which impose the highest penalties to EQ-5D-5L scores (see Appendix E [INSERT LINK TO ONLINE FILE], Table E4, 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, Rice, Robone, & Dias, 2011). We return to this point in the discussion. Nevertheless, we are still able to observe significant variations across unadjusted and DIF-adjusted indices.

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=0.403) and -0.010 (p=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=0.031) based on self-reports to 0.095 (p=0.001) based on DIF-adjustments, while for the medium and low education groups the difference in indices increased from 0.042 (p=0.067) to 0.097 (p<0.001) 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 based on Walters and Brazier (2005).

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=0.051) to 0.155 post-adjustment (above the MID, p<0.001). The difference between Australian-born respondents and those born in Asia increased substantially from 0.156 (p<0.001) to 0.210 (p=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=0.006) to 0.098 (p=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=0.045) to a difference that was not significant (0.055, p=0.266).

Variations across indices for people who were married/de facto and those who were divorced/widowed decreased from 0.086 ($p<0.001$) to 0.063 ($p=0.063$). 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 ($p=0.037$) to 0.123 ($p<0.001$), and from -0.028 ($p=0.406$) to 0.060 ($p=0.100$), 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 ($p=0.067$) to 0.256 ($p<0.001$) for the employed and unemployed, from 0.171 to 0.273 (p for both < 0.001) for the employed and individuals NILF or retired, while decreasing from 0.094 ($p=0.036$) to 0.017 ($p=0.539$) 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 including a covariate for female gender (i.e. the covariate which led to a failure of VE) in the vignette equation of the HOPIT model. Full details of this methodology and findings are described in Appendix D [INSERT LINK TO ONLINE FILE] (results are tabulated in Table E5 of Appendix E); but in general, the robustness checks did not affect our overall conclusions regarding significant differences across subgroups. Moreover, our 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.

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 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, Black, Hollingsworth, & Lorgelly, 2017). 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, Kavetsos, & Tsuchiya, 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- 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 the vignettes were shown, 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, the vignette characters were gender-matched to respondents, and descriptions avoided mention of diseases which could be dependent on age. However we speculate that these measures, designed to make the vignettes as relatable as possible to a diverse audience, contributed to the failure of VE for our sample at large, since they increased heterogeneity in the interpretations of vignettes across respondents.

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. As Kapteyn et al. (2011 p.20) states, "for vignette equivalence to hold, a description has to be complete, minimizing room for different interpretations by different respondents". Thus, by 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 (e.g. specifying a fixed age and gender, which do not vary across respondents), we may reduce the potential for ambiguity in vignette interpretations. However, this may make it more difficult for respondents to use consistent scales across vignette and self-evaluations, and therefore violate RC. Further qualitative and exploratory work is needed in order to understand the trade-off between RC and VE, and whether an appropriate balance can be achieved.

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). The 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.

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. 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.

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*(5), 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. [Article]. *Journal of the Royal Statistical Society Series A: Statistics in Society, 174*, 639-664.

Bago d'Uva, T., O'Donnell, O., & van Doorslaer, E. (2008). Differential health reporting by education level and its impact on the measurement of health inequalities among older Europeans. [Article]. *International Journal of Epidemiology, 37*(6), 1375-1383.

Bago d'Uva, T., van Doorslaer, E., Lindeboom, M., & O'Donnell, O. (2008). Does reporting heterogeneity bias the measurement of health disparities? [Article]. *Health Economics, 17*(3), 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*(4), 875-906.

Bertoni, M. (2015). Hungry today, unhappy tomorrow? Childhood hunger and subjective wellbeing later in life. *Journal of Health Economics, 40*, 40-53.

Brooks, R. (1996). EuroQol: the current state of play. *Health Policy, 37*(1), 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*(7), 621-635.

Crane, M., Rissel, C., Greaves, S., & Gebel, K. (2016). Correcting bias in self-rated quality of life: an application of anchoring vignettes and ordinal regression models to better understand QoL differences across commuting modes. *Quality of Life Research, 25*(2), 257-266.

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*(8), 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*(4), 708-714.

Dowd, J. B., & Todd, M. (2011). Does self-reported health bias the measurement of health inequalities in U.S. adults? Evidence using anchoring vignettes from the Health and Retirement Study. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences, 66*(4), 478-489.

Euroqol Group. (2014). How to use EQ-5D. Retrieved August 2014, from http://www.euroqol.org/about-eq-5d/how-to-use-eq-5d.html

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*(2), 246.

Grol-Prokopczyk, H., Verdes-Tennant, E., McEniry, M., & Ispány, M. (2015). Promises and Pitfalls of Anchoring Vignettes in Health Survey Research. *Demography, 52*(5), 1703-1728.

Hirve, S., Gomez-Olive, X., Oti, S., Debpuur, C., Juvekar, S., Tollman, S., et al. (2013). Use of anchoring vignettes to evaluate health reporting behavior amongst adults aged 50 years and above in Africa and Asia-testing assumptions. *Global Health Action, 6*.

Hopkins, D. J., & King, G. (2010). Improving anchoring vignettes: Designing surveys to correct interpersonal incomparability. *Public Opinion Quarterly, 74*(2), 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*(4), 616-625.

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

Kapteyn, A., Smith, J. P., & Van Soest, A. (2013). Are Americans really less happy with their incomes? *Review of Income and Wealth, 59*(1), 44-65.
Kapteyn, A., Smith, J. P., Van Soest, A., & Voňková, H. (2011). Anchoring vignettes and response consistency. *RAND Working Paper Series WR-840*.
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*(01), 191-207.
Knott, R. J., Black, N., Hollingsworth, B., & Lorgelly, P. K. (2017). Response - Scale Heterogeneity in the EQ - 5D. *Health economics, 26*(3), 387-394.
Kristensen, N., & Johansson, E. (2008). New evidence on cross-country differences in job satisfaction using anchoring vignettes. *Labour Economics, 15*(1), 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*(10), 2187-2196.
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 Approachesto Enhance Cross-Population Comparability of Survey Results *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications* (pp. 421-432). 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*(3), 287-298.
Peracchi, F., & Rossetti, C. (2012). Heterogeneity in Health Responses and Anchoring Vignettes. *Empirical Economics, 42*(2), 513-538.
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*(3), 703-722.
Prosser, L. A., Stinnett, A. A., Goldman, P. A., Williams, L. W., Hunink, M. G. M., Goldman, L., et al. (2000). Cost-Effectiveness of Cholesterol-Lowering Therapies According to Selected Patient Characteristics. *Annals of Internal Medicine, 132*(10), 769-779.
Salomon, J. A., Patel, A., Neal, B., Glasziou, P., Grobbee, D. E., Chalmers, J., et al. (2011). Comparability of Patient-reported Health Status: Multicountry Analysis of EQ-5D Responses in Patients With Type 2 Diabetes. *Medical Care, 49*(10), 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*(10), 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*(4), 410-420.
Sun, S., Chen, J., Johannesson, M., Kind, P., Xu, L., Zhang, Y., et al. (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*(3), 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*(6), 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*(2), 252-260.

Table 1 – Tests of vignette assumptions

	χ^2 test statistic	*p*-value
Response consistency		
Mobility	15.12	0.300
Self-care	18.31	0.146
Usual activities	8.14	0.835
Pain/discomfort	18.86	0.127
Anxiety/depression	19.44	0.110
Vignette equivalence		
Mobility	100.06	<0.001
Self-care	178.69	<0.001
Usual activities	170.03	<0.001
Pain/discomfort	241.63	<0.001
Anxiety/depression	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). Degrees of freedom = 13 for all dimensions/tests. $p<0.05$ indicates a violation of the relevant assumption.

Table 2 – Vignette equivalence – alternate age groups

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

Note: Age 20-23: n=834; Age35-44: n = 1476; Age 45-54: n=871; Age 55-65: n=914. Degrees of freedom = 8 for all dimensions/tests. $p<0.05$ indicates a violation of vignette equivalence.

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

Note: n=914. Degrees of freedom = 40 for all dimensions/tests. $p<0.05$ indicates a violation of reporting homogeneity.

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/retired)					
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)

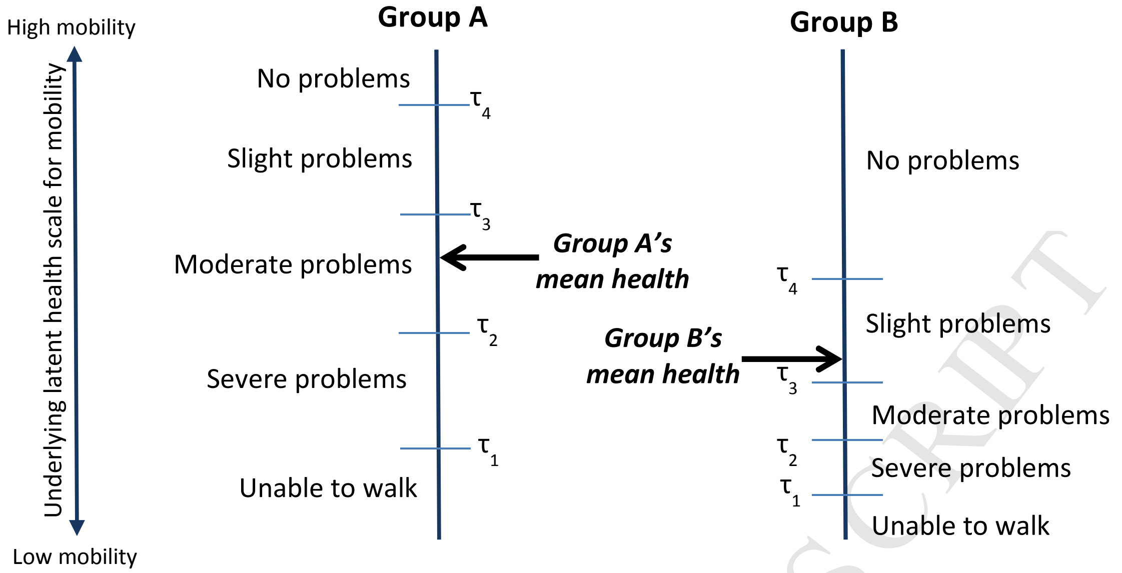
Note: Sample aged 55-65 years (n=917); standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

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

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

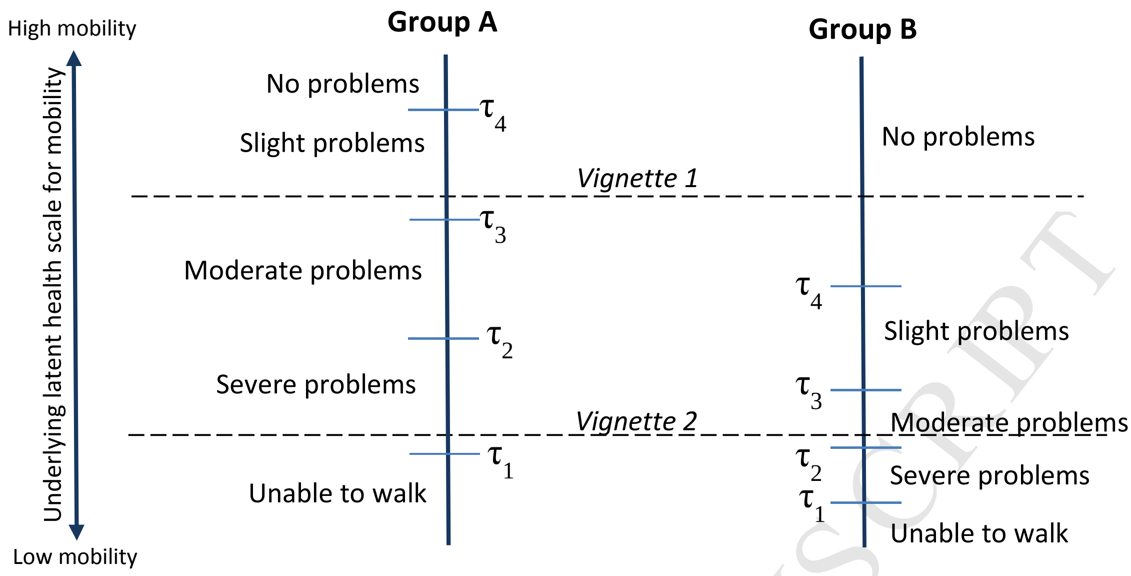
[a]Differences are calculated as the average health score of the first group minus the second group. E.g. Male vs. female is the average health score for males minus that for females.
[b]Calculated from boot-strapped standard errors, 1,000 replications.
[c]Based on a minimally important difference (MID) of 0.074 (Walters and Brazier, 2005).
Note: Based on sample aged 55-65 years (n=917).

Figure 1 – Example of DIF



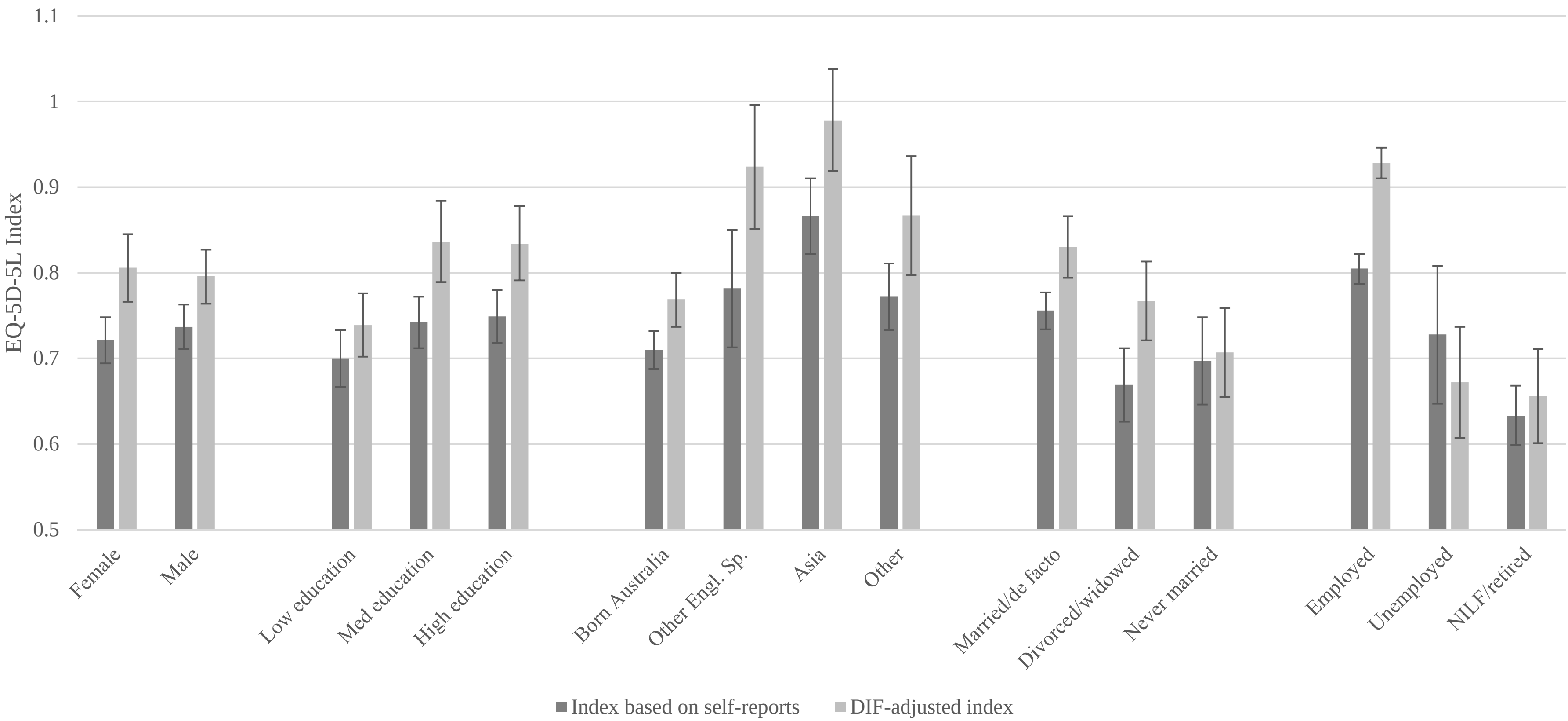
Individuals in Groups A and B are asked to self-report their own level of mobility. How the average individual in each group divides the underlying latent scale into the five response categories is represented by τ_1 to τ_4. DIF is portrayed by variation in the placement of the thresholds across the two groups.

Figure 2 – Logic underlying anchoring vignettes to locate respondent thresholds



Dotted horizontal lines represent the fixed health of each vignette. Group B's ratings are more favourable than Group A's ratings for both vignettes.

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



Unadjusted and DIF-adjusted EQ-5D-5L scores, with associated 95% confidence intervals, for the sample aged 55- 65 (n=917). Confidence intervals are calculated using bootstrapped standard errors.

Highlights:

- Anchoring vignettes can be used to detect differential item functioning (DIF) in the EQ-5D
- The EQ-5D is subject to DIF, which can lead to bias
- The EQ-5D should be used with caution when comparing health across heterogeneous groups