Response-scale heterogeneity in the EQ-5D

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

[a]Centre for Health Economics, Monash University, Clayton VIC 3800, Australia
[b]Division of Health Research, Lancaster University, Lancaster LA1 4YG, United Kingdom

* Corresponding author
Email: rachel.knott@monash.edu
Phone: +61 3 9905 0772
Fax: +61 3 9905 8344

Word count: 2,926
Table count: 0
Figure count: 3

Running Title: Response-scale heterogeneity in the EQ-5D

Keywords: EQ-5D, multi-attribute utility instruments, differential item functioning, reporting heterogeneity, response-scale heterogeneity, preference heterogeneity, anchoring vignettes

Funding: This research was funded by an Australian Research Council Discovery Project [DP110101426].

Conflict of Interest: None

Summary

This paper discusses two types of response-scale heterogeneity which may impact upon the EQ-5D. Response-scale heterogeneity *in reporting* occurs when individuals systematically differ in their use of response scales when responding to self-assessments. This type of heterogeneity is widely observed in relation to other self-assessed measures, but is often overlooked with regard to the EQ-5D. Analogous to this, preference elicitation involving the EQ-5D could be subject to a similar type of heterogeneity, where variations across respondents may occur in the interpretations of the levels (response categories) being valued. This response-scale heterogeneity *in preference elicitation* may differ from variations in preferences for health states which have been observed in the literature. This paper explores what these forms of response-scale heterogeneity may mean for the EQ-5D, and the potential implications for researchers who rely on the instrument as a measure of health and quality of life. We identify situations where they are likely to be problematic and present potential avenues for overcoming these issues.

1. Introduction

Preference-based health-related quality of life (HRQL) measures have grown in popularity, mainly in response to an increasing reliance on economic evaluations to support reimbursement decisions for pharmaceuticals and medical interventions. The most commonly used instrument for measuring HRQL is the EQ-5D, which asks respondents to rate their health in relation to five broad areas or *"domains"*: mobility, self-care, usual activities, pain/discomfort and anxiety/depression (Euroqol Group, 1990). In addition to its inclusion in many economic evaluations, the instrument has recently been used as a performance measure of health-care providers in the UK, and is included in household/health surveys as a measure of population health (Devlin and Krabbe, 2013). The instrument exists in two versions – the original EQ-5D, now renamed the EQ-5D-3L, which includes three levels or response categories (no problems, moderate problems and unable to/extreme problems), and the more recent EQ-5D-5L which contains five response categories for each health domain (no problems, slight problems, moderate problems, severe problems and unable to/extreme problems) (Euroqol Group, 2012).

When the EQ-5D, or any measure using subjective categorical scales, is used as a measure of health, and researchers seek to explain differences in health across heterogeneous patient or population groups (for example, by country, age, gender or socioeconomic status), direct comparisons of self-assessments may be biased if there are systematic differences in the ways that people interpret and use the response scales. For instance, people may rate their health differently on a given categorical scale, either because their true and/or perceived health differs, or because they perceive and use the response scales differently. In the case of the latter, seemingly important differences in self-assessments may falsely be interpreted as differences in underlying health. This phenomenon, a type of reporting heterogeneity which has been referred to as differential item functioning (DIF) (King et al., 2004), differential

reporting behaviour (Rice et al., 2011), or response-scale heterogeneity (Angelini et al., 2014) has been found to exist across a range of subject areas, including self-reported measures of health (e.g. Bago D'Uva et al., 2008b, Grol-Prokopczyk et al., 2011, Kapteyn et al., 2007); political efficacy (King et al., 2004); and job and life satisfaction (Kristensen and Johansson, 2008, Angelini et al., 2014).

Akin to response-scale heterogeneity in reporting, heterogeneity in the interpretations of EQ-5D level descriptions may additionally pose a problem in preference elicitation tasks (i.e. the elicitation of value sets to derive Quality-Adjusted Life Year (QALY) tariffs), since the hypothetical health states that are valued are necessarily constructed using the response categories (levels) of the EQ-5D. Systematic differences in the interpretations of these levels may be masked by, or mistaken for previously observed variation in preferences for health states (e.g. the tendency for older people to value the same health state differently to young people (Dolan, 2000)). This paper highlights the potential issues of response-scale heterogeneity in reporting and preference elicitation, and what they may mean for researchers that use the EQ-5D. Note that while the EQ-5D is the focus of our paper, these issues may apply to other multi-attribute utility instruments such as the SF-6D and Health Utilities Index (HUI). Situations where response-scale heterogeneity is likely to be problematic are identified and potential approaches to overcome these issues in reporting and preference elicitation are outlined.

2. Response-scale heterogeneity in the EQ-5D

2.1. Response-scale heterogeneity in reporting

Two previous studies tested for differences in interpretation and use of EQ-5D response-scales across countries (Salomon et al., 2011, Whynes et al., 2013). They found substantial differences in reporting styles across countries, and that the likelihood of reporting problems

in EQ-5D domains changed considerably after adjusting for DIF (i.e. response-scale heterogeneity).[1] Response-scale heterogeneity has also been found to occur across a range of characteristics (such as age, gender, education and income) in other self-reported measures of health including mobility and pain (which are components of the EQ-5D), sleep, cognition and general health (Bago D'Uva et al., 2008b, Bago d'Uva et al., 2008a, Grol-Prokopczyk et al., 2011). While it has not been formally examined it is likely that response-scale heterogeneity in EQ-5D reporting also extends to subgroups within countries; this may explain at least some of the disparities in EQ-5D responses that have been observed amongst within-country population groups around the world – such as those by age, gender and socio-economic status (Luo et al., 2005, Kind et al., 1999, Burström et al., 2001, Seong et al., 2004, Sun et al., 2011).

An example of response-scale reporting heterogeneity in the pain/discomfort domain of the EQ-5D-3L is illustrated in Figure 1 where the underlying latent scale for perceived pain/discomfort is represented by the vertical line. Assume we wish to compare the health of two groups and respondents are asked to rate their level of pain or discomfort using the response categories "I have no pain or discomfort", "I have moderate pain or discomfort", or "I have extreme pain or discomfort". How each group divides the latent scale into the three response categories is represented by the placement of the inter-category thresholds τ_1 and τ_2. Despite having identical mean levels of latent health (with respect to pain/discomfort) as illustrated by the bold arrows, Group 2 reports moderate pain/discomfort (a utility decrement of 0.123 using UK tariffs, or 0.173 using US tariffs if moving from a state of full health[2], based on general population valuation surveys in each country using the time trade off (TTO) method (Szende et al., 2007)) , whereas Group 1, who are more health optimistic compared to

[1] Whynes et al. (2013) also found significant differences between unadjusted and DIF-adjusted EQ-5D index scores. Salomon et al. (2011) did not examine index values.
[2] i.e., a movement from the EQ-5D-3L health state 11111 to 11121.

Group 2, report no pain/discomfort (which incurs no utility decrement). Researchers would typically be unaware of the location of each group's inter-category thresholds, and may incorrectly conclude that Group 1 is in better health than Group 2. However if the placement of the thresholds could be observed, response-scale reporting heterogeneity would be clearly evident.

<Figure 1 about here>

From this example it is clear how the presence of response-scale heterogeneity may bias conclusions drawn from inter-group comparisons of heterogeneous groups. This bias may be a concern not only in the context of population health surveys but also in economic evaluations alongside clinical trials, where for instance subgroup analyses may be performed to identify cost-effective populations. Figure 2 provides an illustration of how response-scale heterogeneity may present a problem when comparing the health of individuals over time. Here persons A and B are asked to report their mobility before, and sometime after receiving a medical treatment using the EQ-5D-5L. Although the mobility of each individual improves by the same amount on the latent scale post-treatment, Person A reports a health improvement from severe problems to slight problems, whereas Person B, who has a preference toward middle reporting categories, reports no improvement. In the context of a clinical trial, this may mean that an intervention appears more cost-effective if offered to individuals like person A. While these examples describe situations where it is likely to be problematic, it is important to note that response-scale heterogeneity in reporting is unlikely to be an issue for intra-person comparisons in health over time (unless an intervention changes reporting behaviour), or when comparing health across homogenous populations where reporting styles are unlikely to vary.

< Figure 2 about here>

2.2. Response-scale heterogeneity in preference elicitation

When the EQ-5D is used in health technology assessment, health utilities (or HRQL weights) are needed in order to convert patient-reported EQ-5D responses (so called profiles) into QALY weights. Health utilities reflect preferences for health states, and are typically generated by asking members of the public to value a range of health states as defined using the levels (response categories) of the instrument. For example, respondents may be asked to value a hypothetical health state in which they have moderate problems with mobility; moderate problems with self-care, extreme problems doing their usual activities; severe pain/ discomfort; and moderate anxiety/depression (EQ-5D-5L profile of 33543). Individuals completing the preference elicitation task must, in effect, convert these descriptions to an overall state of latent health. If the levels of the EQ-5D mean different things to different respondents, then respondents may not actually be valuing the *same* state of health. To our knowledge this issue has not been explored previously, although the presence of response-scale heterogeneity in preference elicitation may account for some of the variation in preferences for health states observed in the literature (Dolan, 2000). Its existence could also mean that the use of MAUIs and the subsequent QALY calculations in resource allocation decisions may not be giving a true indication of the actual benefit that would be derived, for example, from competing healthcare interventions.

3. Overcoming response-scale heterogeneity

3.1. Response-scale heterogeneity in reporting

A general approach for detecting group differences in reporting styles is to use more 'objective' measures of health, such as biomarkers or more detailed health instruments, to separate differences in health from differences in reporting styles (Salomon et al. (2011) and Whynes et al. (2013) made use of detailed health instruments and clinical measures to detect DIF). However, to avoid confounding of unobserved influences the objective measures must adequately capture variation in true latent health for each of the EQ-5D domains. In practice this can be difficult and costly to achieve. Furthermore, the method does not provide a clear approach regarding how to adjust for reporting heterogeneity once identified (Grol-Prokopczyk et al., 2011).

Another potential solution, which has already been used to identify and correct for response-scale heterogeneity in a number of other self-reported measures of health (Grol-Prokopczyk et al., 2011, Bago D'Uva et al., 2008b, Bago d'Uva et al., 2008a, Kapteyn et al., 2007), is the use of anchoring vignettes.[3] The approach has not yet been used to test for response-scale heterogeneity in the EQ-5D, but pilot research has found the method to be feasible in this context (Au and Lorgelly, 2014). It involves the inclusion of at least one, but typically several, brief health descriptions of hypothetical individuals (vignettes) that respondents are asked to rate using the EQ-5D (in addition to their own health). These ratings can reveal what the response categories truly mean for each respondent, provided that certain identifying assumptions hold, namely response consistency and vignette equivalence (outlined below). Since the actual level of health of the people in the vignettes is the same for all respondents, the variation in respondent ratings can be used to identify and correct for response-scale heterogeneity in reporting.

[3] Note that this approach has also been used to correct for reporting heterogeneity in other areas such as politics and government (King et al. 2004), responsiveness of health system administration (Rice et al 2012), work disability (Kapteyn et al. 2007) and job and life satisfaction (Kristensen and Johnansson 2008; Angelini et al. 2004). Anchoring vignettes have been included in a range of population surveys including the World Health Survey, the Health and Retirement Survey (HRS), the English Longitudinal Study of Ageing (ELSA), and the Survey of Health, Ageing and Retirement in Europe (SHARE).

The intuition of the anchoring vignette approach is illustrated in Figure 3, which extends our first example (Figure 1). Groups 1 and 2 are assumed to divide the underlying latent scale for pain/discomfort as before; and respondents from both groups are given three vignettes describing differing levels of pain or discomfort which they are asked to rate using the same underlying response scale they use to rate their own pain/discomfort. An example of a vignette in this instance may be "Alex suffers from back pain every day and is unable to stand or sit for more than half an hour at a time". In the diagram the dotted horizontal lines represent the fixed health of each vignette. When observing the vignette assessments it is evident that compared to Group 1, Group 2 considers vignettes 2 and 3 to be describing situations of more pain/discomfort.

< Figure 3 about here>

Responses to vignette assessments can be used to identify and correct for response-scale heterogeneity in either nonparametric models or more complex parametric models. The nonparametric approach involves the recoding of self-assessments relative to vignette assessments. The parametric approach entails the estimation of inter-category thresholds as a function of respondent characteristics for each EQ-5D domain (King et al., 2004). The estimated parameters of these models can then be used to predict EQ-5D health states that are not biased by response-scale heterogeneity.

The use of vignettes is particularly appealing as they are simple to complete and less expensive than collecting objective measures of health such as biomarkers, and could potentially be included in surveys alongside the EQ-5D. However, the approach is not without limitation, and there are some conflicting findings concerning the validity of the identifying assumptions upon which the approach is hinged; namely response consistency (RC) - that respondents use the response scales in the same way for the vignettes as they do

for self-assessments; and vignette equivalence (VE) - that the health of the individuals in the vignettes is interpreted in the same way and on the same unidimensional scale across respondents. For instance, studies by Van Soest et al. (2011), Angelini et al. (2014), Rice et al. (2011) and Grol-Prokopczyk et al. (2011) find evidence in favour of these assumptions; while Peracchi and Rossetti (2013), Bago d'Uva et al. (2011) and Datta Gupta et al. (2010) do not. Note that while Peracchi and Rossetti (2013) and Bago d'Uva et al. (2011) find evidence against RC and/or VE, neither study dismisses the approach altogether, rather they draw attention to the role of advances in the design of vignettes as a direction for improving the validity of the methodology. Related to this point and in the context of EQ-5D vignette design, Au and Lorgelly (2014) found that the assumption of RC is more likely to hold if respondents are presented with overall health state vignettes as opposed to short vignettes on single domains. While further work is needed in this area, we are of the opinion that the developing vignette methodology offers a promising approach for addressing response-scale heterogeneity in EQ-5D reporting.

An additional benefit of anchoring vignettes is their potential to make adjustments in external samples, i.e. vignettes collected in one dataset can be used to adjust for response-scale heterogeneity in other datasets where vignettes have not been collected. Harris et al. (2015) have recently shown that such an adjustment is possible, using vignettes of self-assessed health applied to the Household, Income and Labour Dynamics of Australia (HILDA) dataset.[4] The idea of overcoming response-scale heterogeneity by applying an external adjustment has obvious appeal in terms of ease and simplicity. Notably, the approach does require the strong assumption that the reporting behaviour of respondents in the sample which were administered vignettes is the same as the sample in which the external adjustment

[4] Note that this approach does require that the covariates used to predict response-scale heterogeneity are available in both samples; these could include routinely collected characteristics such as gender, age, education and ethnicity.

is being made; such an assumption would need to be thoroughly assessed before such a method was implemented in practice. Given response-scale heterogeneity in reporting has been shown to occur systematically in certain groups, there would appear to be merit in exploring this approach further.

Another potential solution for eliminating response-scale heterogeneity in QALY weights is to produce different subsets of tariffs for different respondent groups (for example according to gender, age or education). Sub-group specific QALY weights could then be applied depending on the characteristics of the particular respondent. By allowing utility weights to vary across respondent characteristics (and if enough sub-groups could be accurately identified), response-scale heterogeneity in reporting would be eliminated in the valuation task, as respondent groups would value health states as they interpret them.[5]. It is worth noting that different sets of tariffs are already applied at the country level, e.g. the UK has a different set of tariffs to the US, these are thought to reflect differences in preferences for overall states of health (Oppe et al., 2007). Dolan (2000) and Flynn and Huynh (2015) have suggested this approach within countries to overcome heterogeneity in preferences across age groups. A limitation is that response-scale heterogeneity in reporting cannot be separated from heterogeneity in preferences.[6] This would mean that it would not be possible through this approach alone to adjust for response-scale heterogeneity in health profiles, which provide rich and detailed information on health states (Parkin et al., 2010).

3.2. Response-scale heterogeneity in the preference elicitation task

[5] The approach would rely on the assumption that interpretation and use of the response scales (i.e. the levels of the EQ-5D) are the same amongst all people in a particular sub-group. I.e. the approach would assume that, in each particular sub-group, response-scale use is the same for respondents who complete the EQ-5D self-assessment (to which QALY weights are applied) and respondents who complete the valuation task (from which the QALY weights are derived).

[6] Although other approaches could still be used to detect response-scale heterogeneity in reporting.

How to detect and overcome response-scale heterogeneity in the preference elicitation task is much less clear; and discussion of this has been minimal. The inclusion of anchoring vignettes in the valuation task could make its detection possible. For instance, if anchoring vignettes were able to reveal what EQ-5D levels truly meant to respondents undertaking the valuation activity, the elicited values could be adjusted to correct for response-scale heterogeneity before deriving the QALY tariff. The approach discussed above involving the elicitation of subgroup tariffs could also be used to (theoretically) eliminate response-scale heterogeneity in the preference elicitation task, as well as in reporting.

4. Concluding remarks

This paper highlights the important issues of response-scale heterogeneity in EQ-5D reporting and in preference elicitation. It is hoped that emphasising these issues in the context of the EQ-5D will lead to further discussion and exploration in health instruments in general, so that ultimately these potential sources of bias can be addressed in analyses that use self-reported health measures to make comparisons across heterogeneous groups. We outline a number of methods for overcoming response-scale heterogeneity in EQ-5D reporting, and suggest ways forward in terms of the exploration of response-scale heterogeneity in preference elicitation[7]. It is our opinion that the most promising of the approaches discussed are the application of anchoring vignettes and the construction of alternate sets of tariffs for different population groups. In practice, the latter would be an expensive approach to implement as large samples would be required to appropriately represent each of the distinctive combinations of population groups (e.g. according to age, gender, education, etc). The approach would however eliminate *all* response-scale heterogeneity in the EQ-5D and may be worth considering in future work. Perhaps the most viable method worth pursuing in

[7] Note that due to limitations of space we do not provide an exhaustive review; for instance, Rasch analysis can also be used to identify response-scale heterogeneity.

the immediate future, in terms of both time and cost, is the anchoring vignette approach, particularly if adjustments can be formulated using vignette responses from external samples (if the assumptions required for this approach prove valid).

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., 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.

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.

DATTA GUPTA, N., KRISTENSEN, N. & POZZOLI, D. 2010. External validation of the use of vignettes in cross-country health studies. *Economic Modelling,* 27**,** 854-865.

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.

DOLAN, P. 2000. Effect of age on health state valuations. *Journal of Health services Research & policy,* 5**,** 17-21.

EUROQOL GROUP 1990. EuroQol-a new facility for the measurement of health-related quality of life. *Health Policy,* 16**,** 199-208.

EUROQOL GROUP. 2012. *Interim scoring for the EQ-5D-5L: EQ-5D-5L Crosswalk Index Value Calculator* [Online]. Available: http://www.euroqol.org/news/news/article/interim-scoring-for-the-eq-5d-5l-eq-5d-5l-crosswalk-index-value-calculator.html.

FLYNN, T. & HUYNH, E. 2015. Best-Worst Scaling Profile Case Application: Preferences for Quality of Life in Australia. *In:* LOUVIERE, J., FLYNN, T. & MARLEY, A. (eds.) *Best-Worst Scaling: Theory, Methods and Applications.* Cambridge, UK: Cambridge University Press.

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.

HARRIS, M. N., KNOTT, R. J., LORGELLY, P. K. & RICE, N. 2015. Using external vignettes to correct for reporting heterogeneity. *Mimeo*.

JANSSEN, B. M., OPPE, M., VERSTEEGH, M. M. & STOLK, E. A. 2013. Introducing the composite time trade-off: a test of feasibility and face validity. *The European Journal of Health Economics,* 14**,** 5-13.

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.

KIND, P., HARDMAN, G. & MACRAN, S. 1999. *UK population norms for EQ-5D*, Centre for Health Economics, University of York UK.

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.

KRISTENSEN, N. & JOHANSSON, E. 2008. New evidence on cross-country differences in job satisfaction using anchoring vignettes. *Labour Economics,* 15**,** 96-117.

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.

OPPE, M., DEVLIN, N. J. & SZENDE, A. 2007. EQ-5D value sets: inventory, comparative review and user guide.

PARKIN, D., RICE, N. & DEVLIN, N. 2010. Statistical analysis of EQ-5D profiles: does the use of value sets bias inference? *Medical Decision Making,* 30**,** 556-565.

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.

RICE, N., ROBONE, S. & SMITH, P. 2011. Analysis of the validity of the vignette approach to correct for heterogeneity in reporting health system responsiveness. *The European Journal of Health Economics,* 12**,** 141-162.

RICE, N., ROBONE, S. & SMITH, P. C. 2012. Vignettes and health systems responsiveness in cross‐country comparative analyses. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 175**,** 337-369.

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.

SEONG, S. S., CHOI, C. B., SUNG, Y. K., PARK, Y. W., LEE, H. S., UHM, W. S., KIM, T. W., JUN, J. B., YOO, D. H. & LEE, O. Y. 2004. Health-related quality of life using EQ-5D in Koreans. *The Journal of the Korean Rheumatism Association,* 11**,** 254-262.

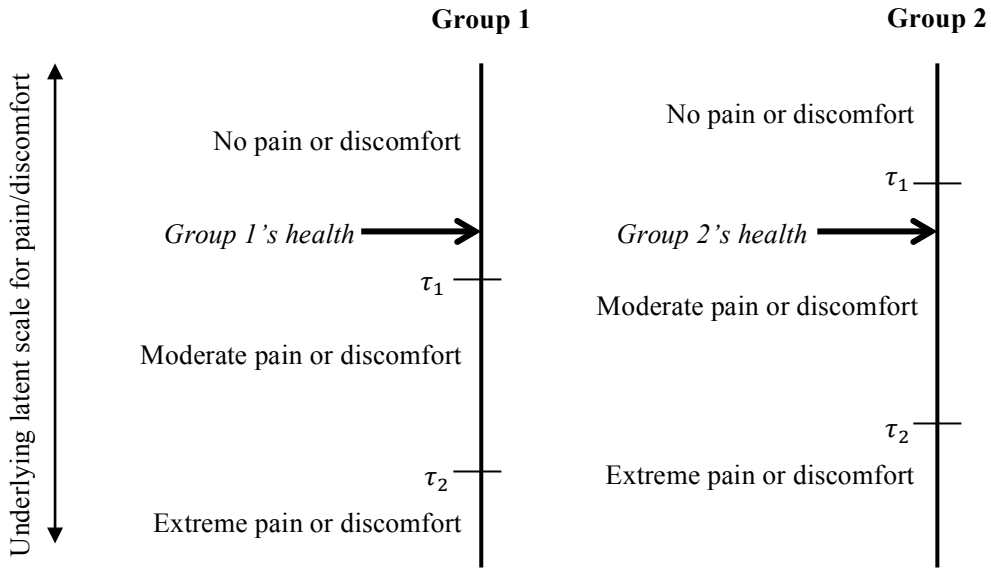
SUN, S., CHEN, J., JOHANNESSON, M., KIND, P., XU, L., ZHANG, Y. & BURSTRÖM, K. 2011. Population health status in China: EQ-5D results, by age, sex and socio-economic status, from the National Health Services Survey 2008. *Quality of Life Research,* 20**,** 309-320.

SZENDE, A., OPPE, M. & DEVLIN, N. 2007. EQ-5D value sets: Inventory. *Comparative Review, and User Guide: Spinger*.

VAN SOEST, A., DELANEY, L., HARMON, C., KAPTEYN, A. & SMITH, J. P. 2011. Validating the use of anchoring vignettes for the correction of response scale differences in subjective questions. *Journal of the Royal Statistical Society. Series A: Statistics in Society,* 174**,** 575-595.

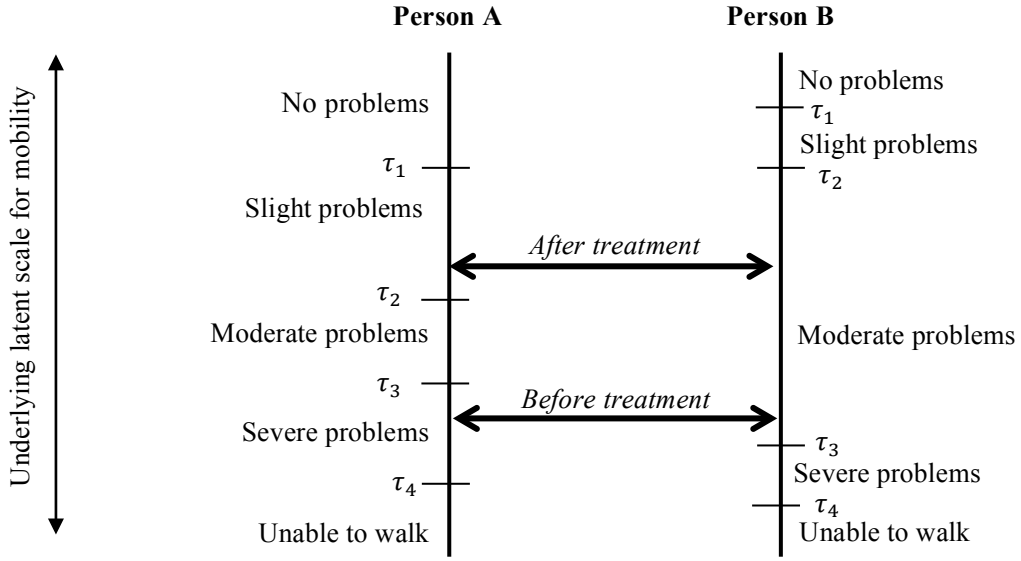
WHYNES, D. K., SPRIGG, N., SELBY, J., BERGE, E., BATH, P. M. & INVESTIGATORS, E. 2013. Testing for Differential Item Functioning within the EQ-5D. *Medical Decision Making,* 33**,** 252-260.

Figure 1 – Response-scale heterogeneity in the EQ-5D-3L pain/discomfort domain (Example 1)



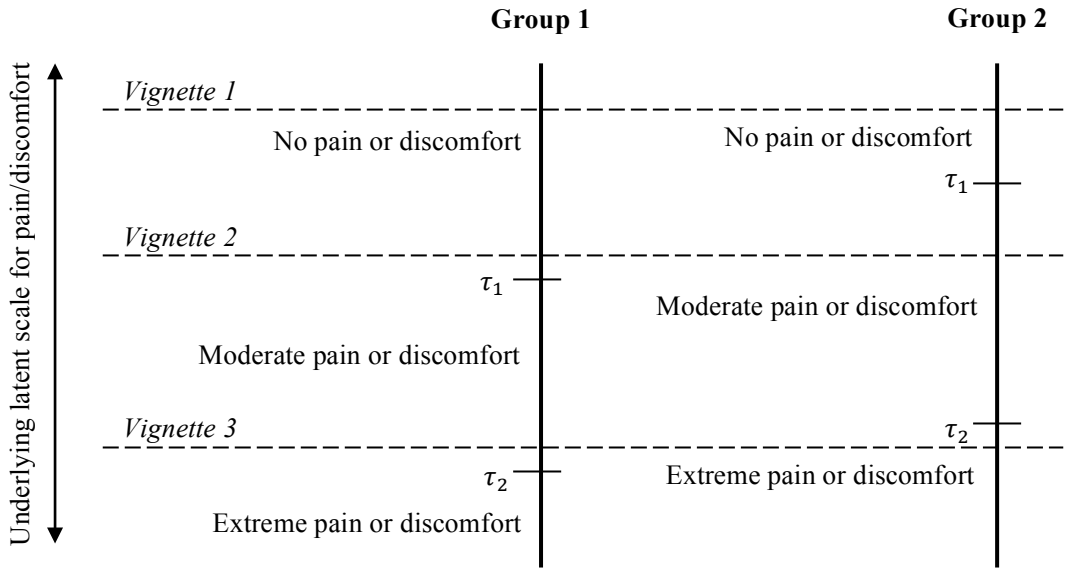
Groups 1 and 2 are asked to report their level of pain/discomfort using the response categories "I have no pain or discomfort", "I have moderate pain or discomfort" and "I have extreme pain or discomfort". Each group divides the underlying latent scale for pain/discomfort (vertical line) into categories according to the placement of the inter-category thresholds (τ_1) and (τ_2). Although mean levels of latent health (in terms of pain/discomfort) are the same for each group, as shown by the bold arrows, Group 1 reports no pain or discomfort whereas group 2 reports moderate pain or discomfort.

Figure 2 – Response-scale heterogeneity in the EQ-5D-5L mobility domain before and after an intervention (Example 2)



Persons A and B are asked to report their mobility before and sometime after receiving a medical treatment using the EQ-5D-5L. Although the mobility of each individual improves by the same amount on the latent scale post treatment, Person A reports a mobility improvement from severe problems to slight problems, whereas Person B reports no improvement in mobility.

Figure 3 – Anchoring vignettes in the EQ-5D-3L pain/discomfort domain (Example 1 continued)



Respondents from Groups 1 and 2 are given three vignettes describing differing levels of pain or discomfort which they are asked to rate using the EQ-5D-3L. The fixed health of each vignette is represented by the dotted horizontal lines. Group 1 considers vignettes 1 and 2 to be describing situations of no pain/discomfort and vignette 3 to be a situation of moderate pain/discomfort. Group 2 considers vignette 1 to describe no pain/discomfort; vignette 2 to describe moderate pain/discomfort; and vignette 3 to describe extreme pain/discomfort.