Preference Weighting of Health State Values: What Difference Does It Make, and Why?

Admassu N. Lamu[1]*, Thor Gamst-Klaussen[1], Jan Abel Olsen[1]

[1]Department of Community Medicine, University of Tromsø,

N-9037 Tromsø, Norway

***Corresponding author:** Admassu N. Lamu

admassu.n.lamu@uit.no

Abstract

Background: Most patient-reported outcome measures apply a simple summary score to assess health related quality of life, whereby equal weight is normally assigned to each item. In the generic preference-based instruments, utility weighting is essential whereby health state values are estimated through preference elicitation and complex algorithms. Objective: This paper examines the extent to which preference-weighted value sets differ from unweighted values in the EQ-5D-5L and 15D instruments, based on a comprehensive dataset from six OECD countries, each with a representative healthy sample and seven disease groups (N=7933). Methods: Construct validities were examined. The level of agreement between preference-weighted vs. unweighted values was also assessed using intra-class correlation coefficient (ICC), Bland-Altman plots, and reduced major axis (RMA) regression. Results: The performance of preference-weighted and unweighted measures were comparable with regard to convergent and known-group validities for each instrument. While unweighted EQ-5D-5L values differ considerably from the preference-weighted values at individual level, the discrepancy is minimal at the group level with mean difference of 0.023. The ICC (0.96) and Bland-Altman plot also suggest strong overall agreement. For the 15D, both ICC (0.99) and Bland-Altman plot revealed almost perfect agreement, with a negligible mean difference of -0.001. Results from RMA regression also showed small bias. Conclusions: Overall, preference weighting has minimal effect if the unweighted values are anchored on the same scale as the preference-weighted value sets, at least at the group-level.

Keywords: Health related quality of life; preference weighting; EQ-5D-5L; 15D.

1. Introduction

A wide range of instruments have been developed to measure patient reported outcomes, often by use of a summary score to indicate the degree of disease severity [1]. The majority of these instruments assign equal weight to each dimension or item included, i.e. every health dimension and each level change are assumed to have equal importance. Furthermore, these instruments do not account for how people value a health state improvement relative to how they value lifetime gains.

Generic preference-based instruments are different. They were designed to enable comparisons of the effectiveness of competing health care programmes in economic evaluations [2, 3]. Since effectiveness can be in terms of both improved health and prolonged life, the health-related quality of life (HRQoL) gains are made commensurable with lifetime gains, using a scale that account for people's trade-offs between quality and quantity of life. Furthermore, reflecting economists' attention to the preferences of affected parties, these instruments also seek to account for importance weighting of the included health dimensions. The distinct features of these preference-based instruments are that they: i) use a generic health state descriptive system designed to apply across all health conditions, and; ii) provide an indirect means of obtaining preference weights. Hence, respondents are assigned a health state value based on their responses to a health state questionnaire, and pre-specified preference weights obtained from other populations are then applied [4]. The focus on utility represents a key element, in that the class of cost-effectiveness analysis based on these instruments are referred to with a specific term; cost-utility analyses (CUA).

The most widely used health state utility instrument is the EQ-5D, followed by SF-6D, the HUI, and the 15D. Together, these four instruments are found in around 95% of applied cost-utility studies [5]. Further, a review of 1,663 studies using preference-based instruments published between 2005 and 2010 found that the EQ-5D had been applied in 63% of these studies [6]. In addition to their different descriptive systems, these instruments apply different preference elicitation methods: the visual analogue scale (VAS), or the choice based methods of time-trade-off (TTO), standard gamble (SG), and discrete choice experiments (DCE). Furthermore, different scoring algorithms are used. Consequently, different instruments produce different preference weights [7, 8].

Several researchers have questioned the complex algorithms used to create preference weights [9-11]. Richardson, Iezzi, and Khan [12] suggest that differences in preference weights are primarily via their effect upon the measurement scales. Although each preference-based measure was developed on a unit scale of 0 to 1, their actual scales differ: the original English value set for the EQ-5D has a scale length of 1.594 (i.e. -0.594 to 1), while the SF-6D has a scale length of 0.797 [6]. The aim of this paper is to examine what difference it makes to assign preference-weighted values to health states, as compared to the unweighted values obtained when summary scores are converted onto a [0 – 1] scale. Given that some preference-based instruments include negative values, reflecting that the most inferior health states are considered worse than being dead, parts of the discrepancy between preference-weighted vs. unweighted values are explained by scale-differences. Hence, a key issue is to make scale-adjusted comparisons, in order to determine how much of the observed discrepancy is due to scale length differences, and how much is attributable to the importance weighting of health dimensions.

This paper examines two preference-based instruments, EQ-5D-5L and 15D, which are contrasting in terms of both their descriptive system and valuation methods. The EQ-5D-5L has the most condensed descriptive system, including only five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression [13]. In the construction of the new EQ-5D-5L, the original dimensional structure was retained, but now includes five levels of severity (no problems, slight problems, moderate problems, severe problems, and unable to/extreme problems) [14]. The 15D describes health along 15 dimensions (mobility, vision, hearing, breathing, sleep, eating, speech, bladder/bowel function, usual activities, mental function, discomfort/pain, depression, distress, vitality, sexual activity), each with five levels, giving a combination of over 30.5 billion (=5^{15}) possible health states [15].

As for valuation methods, in the 15D subjects were asked to rank the dimensions and the levels within each dimension according to their relative importance using a 0 to100 VAS scale, where 100 was given to the most important dimension or level, and 0 was assigned if a dimension or level was not considered important at all [15]. The EQ-5D-5L tariff considered here is the latest version, based on an English population sample. It applies a combination of TTO and DCE tasks, which makes explicit trade-offs between quality and quantity of life, with scales below zero [16].

2. Data and Methods

2.1 Data

Data was obtained from the Multi-Instrument Comparison (MIC) study, which is based on an online survey administered in Australia, Canada, Germany, Norway, UK and the USA by a global panel company, CINT Australia Pty Ltd [17]. The personal and medical details recorded by the panel company were used to recruit individuals from a 'healthy group' (N=1760) and from seven major chronic disease groups (N=6173). Quotas on age, gender and education were used to obtain a demographically representative sample of 'healthy' respondents, defined by the absence of chronic disease and a VAS score of at least 70 on overall health. Quotas were also applied to obtain a target number of respondents in each disease group: arthritis, asthma, cancer, depression, diabetes, hearing loss, and heart problems.

In addition to the MIC dataset, the full set of EQ-5D-5L health states ($N=5^5=3125$) were employed to explore the degree of agreement between preference-weighted and unweighted values. For the 15D, however, all analyses were based on the MIC dataset as it is problematic to use the 30.5 billion full set of 15D health states. For the purpose of comparing preference-weighted and unweighted values in both the EQ-5D-5L and 15D in terms of construct validity, four variables were considered: two variables (VAS and standard of living) correspond to the full sample (N=7933); and the other two (diabetes 39, D-39, and Kessler Psychological Distress Scale, K10) were taken from the sub-sample of 'disease groups'. The D-39 and K10 were chosen since they were relatively more interrelated with both EQ-5D-5L and 15D dimensions.

2.2 Preference-Weighted Scoring Approach for the EQ-5D-5L and 15D

2.2.1 The EQ-5D-5L

Health states defined by the EQ-5D-5L may eventually be converted to a single summary index by applying scores from a standard set of values (preferences) derived from general population samples [18]. In the current paper, the value set for EQ-5D-5L is derived from the stated preference data of 996 members of the English general public, where a hybrid model combining a composite TTO (cTTO) approach and DCE tasks were used for its direct elicitation [16]. The minimum value for the worst health state ('the pits') is -0.281, giving a scale length of 1.281 (i.e. from -0.281 to 1).

2.2.2 The 15D

The 15D tariff was generated using a set of preference weights elicited from several representative samples of the Finnish adult population [15]. Respondents were asked to assign the relative importance for 15D dimensions on a 0 to 100 scale, where 100 was given to the most important dimension. Then the importance of all other dimensions were assessed in relation to this most important dimension. Similarly, importance weights for levels within-dimension were produced on a 0 to 100 scale, where the most desirable level (level-1) assigned 100 and the desirability of all other levels were assessed in relation to level-1. In addition to the five levels, the states of 'unconscious' and 'dead' were also valued for each dimension. The preference weights are scaled on a [0 – 1] range, where 0 representing 'dead' and 1 'no problems on any dimension', and with no health state worse than being dead. The weights were obtained by use of a rating scale (i.e. VAS) and then combined using a simple additive model. Hence, the 15D value set is not based on preferences that reflect for trade-offs between quality vs. quantity of life gains.

2.3 The Unweighted Scoring Approach

Based on the instruments' summary scores, unweighted health state values are developed, with each dimension assigned equal importance and each level change assigned the same weight. First, item scores are set equal to the rank order of the reverse coded response (so that higher values correspond with better health), and summed to obtain a summary score, X_i, for each health state i. Then, X_i is constrained to the range [0 – 1] to obtain unweighted values, V_i, using a unity based normalization equation as follows:

$$V_i = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \qquad (1)$$

where X_{min} and X_{max} are the summary scores obtained when the response to every item of the instrument is at its minimum (worst) and maximum (best) level respectively. For instance, because of reverse coding, a health state 11232 on the EQ-5D-5L becomes 55434, and hence X_i for this health state is 21 (i.e. 5+5+4+3+4). Again, because of reverse coding of the worst health state 55555 into 11111 and vice versa, X_{min} is 5 (1+1+1+1+1) and X_{max} is 25 (5+5+5+5+5). Therefore, the unweighted value for the health state 11232 on [0 – 1] scale is 0.80; i.e., (21-5)/(25-5). According to this scale, the unweighted EQ-5D-5L has 20 possible values with an interval of 0.05 (=1/20) between each successive values, whilst unweighted 15D has 60 different possible values with an interval of 0.0167 (=1/60).

Equation (1) gives a simple unweighted value on a [0 – 1] scale without adjustment to the scale of the preference-weighted tariffs. However, to enable comparisons on the same scale, we perform a simple linear transformation onto the same scale as the weighted utility range, i.e. [-0.281 – 1] for EQ-5D-5L. This is achieved by using min-max normalization approach described by Han et al. [19], which preserves the relationships among the original data values.

$$V_i^{'} = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} * \left(X_{\max}^{'} - X_{\min}^{'}\right) + X_{\min}^{'} = V_i(Range) + X_{\min}^{'} \qquad (2)$$

where X'_{min} and X'_{max} are the minimum and maximum possible values on the preference-weighted tariffs, and V'_i represents the unweighted values on the same scale as the preference-weighted scale.

For instance, the algorithm for computing the V'_i for the above health state of EQ-5D-5L (11232) on the [-0.281 – 1] scale is: $V'_i = V_i*(\text{Range}) + X'_{min}$ = 0.80*[1.00 - (-0.281)] + (-0.281) = 0.744. With this linearly transformed scale, the interval between successive values of V'_i becomes 0.064 (=1.281/20) for the EQ-5D-5L. For the 15D, the preference-weighted tariff is anchored on a [0 – 1] scale which coincides with the scale in equation (1) and hence no linear transformation is needed.

Both the V_i and V'_i obtained in equations (1) and (2) refer to equally weighted (or unweighted) values. However, while equation (1) represents a [0 – 1] scale, the V'_i in equation (2) accounts for a scale range including negative values. Consequently, preferences for the trade-off between gains in quality vs. quantity of life (the scaling issue) is indirectly reflected in it, and any difference from its preference-weighted counterpart is only the non-equal importance weighting depending on which health dimensions a given quality gain will occur. Hence, when comparing preference-weighted vs. unweighted values, equation (2) adjusts for the part of these discrepancies that reflect scale-differences.

2.4 Statistical Analysis

2.4.1 Convergent Validity

To determine the extent to which the preference-weighted and unweighted values are related to other measures of similar construct, convergent validity was examined by comparing them to the scores reported on the visual analogue scale (VAS) for the total sample (N=7933), and to the Diabetes-39 sub-sample (N=924) using Spearman rank order correlations. The Diabetes-39 (D-39) is a disease-specific

instrument for diabetes patients with 39-items, each with 7-response level ranging from 1 (not affected at all) to 7 (extremely affected) [20]. It covers five dimensions: energy and mobility (15 items), diabetes control (12 items), anxiety and worry (4 items), social burden (5 items), and sexual functioning (3 items). Each attribute was reverse-coded and the total score on each domain was linearly transformed to a [0 – 1] scale; 0 indicating the worst; and 1 the best possible health state. Convergent validity with the D-39 sub-scales were also assessed. We would expect strong correlations between VAS and the preference-weighted as well as the unweighted values. As for D-39, we expect high correlation with 'energy and mobility' as well as 'anxiety and worry' subscales (since both EQ-5D-5L and 15D dimensions cover these sub-scales).

2.4.2 Known Group Validity

A known group validity was tested to examine the discriminative validity of the preference-weighted and unweighted values for each instrument. Current standard of living (very good, good, poor, and very poor) was used as a reference for the whole sample. In addition, the Kessler Psychological Distress Scale (K10) was used as an anchor in the depression patient group (N=917). Following Jorm et al. [21], the K10 is re-categorized into four severity levels: 'likely to be well' (10 – 19), 'mild' (20 – 29), 'moderate' (30 – 39), and 'severe' (40 – 50). Subjects with poorer health status and standard of living were hypothesized to have lower scores. Kruskal-Wallis test and relative efficiency (RE) were employed to explore the known-group validity of preference-weighted and unweighted values for both EQ-5D-5L and 15D. The RE statistic is defined as the ratio of either chi-squared statistics or squared t-statistics [11]. Here RE is given as the ratio of chi-squared (χ^2) of preference-weighted and unweighted values. An RE value greater than 1 implies that the preference-weighted tariff has more power in discriminating between meaningfully different groups, and the converse is true for an RE value less than 1.

2.4.3 Level of Agreement

The degree of agreement between preference-weighted and unweighted values was assessed based on the intra-class correlation coefficient (ICC) [22], Bland-Altman plots [23], and reduced major axis (RMA) regression for each instrument. The ICC was constructed based on a two-way mixed effects model with absolute agreement, and a single measure of ICC was calculated. The Bland-Altman analysis involves computing the mean and the difference between measurement methods for each subject in the sample. It reports the population mean difference between the two methods, and the 95% limits of agreement that provide a limit within which 95% of the variability between the methods will lie. RMA is used to

detect bias between two measures [24]. Its slope provides an estimate of the amount of systematic bias. The results of RMA are reported graphically to visualize how the level of agreement between preference-weighted vs. unweighted values differ with scale-length. All statistical analyses were conducted using Stata® ver. 14.1 (Stata Corp LP, College Station, Texas, USA).

3. Results

3.1 Convergent and Known Group Validity

There is evidence of convergent validity of preference-weighted and unweighted values for each instrument (EQ-5D-5L and 15D) with both VAS and D-39 scores (Table 1). The rank correlation between the VAS and the preference-weighted and unweighted measures of each instrument was high (0.60 and higher). Similarly, all Spearman rank order coefficients for the preference-weighted and unweighted values with the five D-39 domains were significant ($p<0.001$). Correlations were highest for 'energy and mobility' domain (0.70 and higher), as expected. Relatively high correlations were also found with 'anxiety and worry' dimension. The unweighted measures demonstrate similar performance in terms of convergent validity compared with the preference-weighted scores both in the EQ-5D-5L and in the 15D instruments.

[Insert Table 1 about here]

Both preference-weighted and unweighted measures of EQ-5D-5L and 15D give evidence of known group validity in detecting significant ($p < 0.001$) differences between the known group variables (standard of living and depression, K10) (Table 2). The preference-weighted EQ-5D-5L appears to be more effective in discriminating both groups with RE significantly more than 1; i.e. (RE = 1.05; 95% CI: 1.030, 1.071) when standard of living is used, and (RE=1.14; 95% CI: 1.055, 1.227) when K10 is applied. Preference-weighted 15D appears to have less discriminating power as compared to its unweighted counterpart in both comparison groups with RE significantly less than 1.00 (Table 2).

[Insert Table 2 about here]

3.2 Agreement between Preference-Weighted and Unweighted Values

The Spearman correlation between preference-weighted and unweighted EQ-5D-5L is very high, indicating a good degree of association (Table 3). The scale of the instrument, however, influences the level of agreement. For instance, our results reveal a substantial agreement for EQ-5D-5L [ICC = 0.96 (95% CI: 0.931, 0.969)] when the preference-weighted and unweighted values are given on the same scale. If unweighted values are anchored on the [0 – 1] scale, the agreement is weaker, particularly when the full set of health states (5^5=3125) is used, instead of the MIC dataset; i.e. ICC rises from 0.76 to 0.92 with adjustment in the scale of unweighted values (results for the full set of health states are not reported here).

Similarly, the Bland-Altman plots shown in Figure 1 suggest that the preference-weighted and unweighted values of EQ-5D-5L has a high level of agreement at the group level. The mean difference is similar (about 0.02) when the MIC data is considered, irrespective of whether unweighted values are adjusted to the preference-weighted scale or not. When we consider the 3,125 possible health state combinations in the EQ-5D-5L descriptive system, the mean difference is 0.03 (95% CI: 0.029, 0.035) for the scale adjusted, and -0.11 (95% CI: -0.112, -0.105) for the unadjusted one. Thus, the mean bias is more than tripled if we do not adjust for the difference in the scales. The RMA regression depicted in Figure 2 demonstrate similar results, with slope closer to 1 and intercept closer to 0 when scale adjusted unweighted values are employed.

Despite a small overall mean difference between preference-weighted and unweighted EQ-5D-5L, a large inter-individual difference is evident. The lower and upper 95% limits of agreement for the EQ-5D-5L is -0.085 (95% CI: -0.087, -0.083) and 0.131 (95% CI: 0.129, 0.133), respectively. The corresponding limits of agreement for the full set of health states is -0.124 (95% CI -0.129, -0.119) and 0.188 (95% CI: 0.183, 0.193). The Bland-Altman plot for the EQ-5D-5L (Figure 1 (a)) indicates some systematic variation at the lower end of the scale, which is likely due to the fact that there are relatively large utility decrements associated with levels 4 and 5 on the ‘pain/discomfort’ and ‘anxiety/depression’ dimensions. The number of observations outside these limits of agreement is 7.41% for EQ-5D-5L.

[Insert Table 3 & Figure 1 about here]

As for the 15D, the Spearman rank correlation is very high (ρ = 0.99), and so is the agreement between preference-weighted and unweighted distributions. The ICC (0.99; p < 0.001), which measures the absolute agreement, suggests a nearly perfect agreement. In a pairwise comparison between preference-weighted and unweighted 15D, the mean difference is negligible at the group level (-0.001 with 95% CI: -0.002, -0.001). The 95% limits of agreement depicted in Figure 1 is -0.038 (95% CI: -0.039, -0.037) to 0.036 (95% CI: 0.035, 0.037)), indicating small difference even at the individual level (Table 3). Only 5.9% observations lie outside these limits of agreement. The RMA regression results reported in Figure 2 also reveal little bias between preference-weighted and unweighted 15D.

[Insert Figure 2 about here]

4. Discussion

We have examined the effect of preference weighting in two instruments; the EQ-5D-5L and the 15D, in terms of validity and the level of agreement. The results reveal that the preference-weighted and the unweighted measures for each instrument were strongly correlated with the VAS and the D-39, and each measure was able to discriminate differences between known groups. However, whilst the unweighted EQ-5D-5L revealed slightly poor known group validity, the unweighted 15D showed better performance as compared to the preference-weighted version. With respect to agreement between preference-weighted and unweighted values, a simple comparison of the mean values in the EQ-5D-5L for the whole population generally reveals small discrepancy. While the mean difference is negligible at the group level, the individual difference between weighted and unweighted values is modest in the 15D. However, the most widely used instrument (EQ-5D-5L) showed a considerable discrepancy at an individual level.

Previous studies suggest that greater reliability and validity might be achieved by simply using unweighted values rather than the increasingly complex algorithm of utility weights [4, 25, 26]. For instance, Prieto and Sacristán [10] argued that the weighting system in the preference-based instruments does not indicate a substantial difference in the final score from that of unweighted values for EQ-5D-3L. Similarly, Wilke et al. [11] found no difference in sensitivity to change between weighted and unweighted values, although the weighted values better discriminate between disease groups, and unweighted values provide a greater test-retest reliability for the EQ-5D-3L and the HUI-3. In similar

vein, our results reveal that preference weighting produces a small difference when the unweighted values are adjusted to the same scale as the preference-weighted values in EQ-5D-5L, at least at the group level.

While the scale length reflects preferences over quality vs. quantity, there are two different theoretical reasons to expect discrepancies between preference-based values and the simplified scale adjusted values presented in this paper. Firstly, there is nothing to suggest why people should have identical preference weights on qualitatively different dimensions. The study on which the English EQ-5D-5L value set is based shows the last two dimensions (pain/discomfort, anxiety/depression) have higher preference weights than the first three dimensions (mobility, self-care, usual activity). The sum of the first three's weights is about the same as the sum of the last two's weights [16]. Secondly, health state utility instruments are descriptive systems as opposed to a Likert scale with identical intervals between numbers. Hence, the utility drops from one level to the next level down reflect the severity differences associated with the words used. The English value set for EQ-5D-5L reveals clear non-linearities along all dimensions with around half of the total utility decrement occurring between levels 3 and 4 [16].

It is interesting to compare preference-weighted and unweighted values at the individual level, since the theoretical arguments for the use of preference weights are technically valid at the individual level [3]. Our result indicates a clear discrepancy in EQ-5D-5L at the individual level with the width of the 95% limits of agreement equal to 0.216 for the MIC dataset and 0.312 for the full set of health states. However, for population mean, the adjusted unweighted values appear to give similar results to the preference-weighted tariffs (with mean difference closer to 0.02). This difference is much lower than the clinically importance difference (0.074) reported for EQ-5D-3L [27].

The range of the instrument scale is crucial in the comparison between preference-weighted vs. unweighted values. The preference-based HRQoL instruments were developed with the intention that utilities are measured on a cardinal scale of [0 – 1], where 0.00 represents being dead and 1.00 perfect health. States worse than dead are accounted for by assigning negative values. For example, the effective ranges for EQ-5D-5L is 1.281 (-0.281 to 1). However, the unweighted scale based on normalizing the summary scores can never go below zero. Obviously, this scale difference accounts mainly for the difference between preference-weighted and unweighted values. For instance, the level of agreement between preference-weighted and unweighted values rises substantially after adjusting

for the scale differences (i.e. ICC rises from 0.76 to 0.92) when the full set of health states are considered. However, the corresponding change in ICC is quite small (Table 3) with the MIC dataset that comprises health states, which people actually experience. This is mainly because the majority of respondents (over 80%) did not experience health state combinations with high severity level (level 4/5 on any dimension). In general, the differences between preference-weighted and unweighted values arise primarily due to scale effect brought up by the methodological approach used to construct preference weights [10]. Preference weights also determine the measurement scale of an instrument [12], which has an impact on the calculation of QALY and hence the results of cost-utility analyses.

With regard to 15D, our results reveal only negligible difference between a preference-weighted and an unweighted value. The overall mean difference is close to zero (-0.001). This mean difference is by far lower than the generic minimum important changes (0.015) reported for the 15D scores [28]. One possible explanation could be related to similarity of the scale range. The worst possible health state (the 'pits') has a value of zero for both the preference weighted and unweighted scale. Furthermore, the 15D has many dimensions that allows for a large number of health state combination (5^{15}), which leads to the compression of weights [12]. Thus, in the absence of scale length difference, preference weighting that involves mere relative importance brings small difference. Note that unlike the choice based techniques, the rating scale (VAS) is not a utility instrument because respondents are not requested to sacrifice anything (life years or risk of death). Therefore, given such minimal effect of assigning different importance weighting to the various levels of 15D dimensions, a simple linear transformation of its summary scores, equation (1), might suffice or even be superior to preference-weighted tariffs.

This study highlights the implications of scale differences arising from different preference weighting algorithms and valuation techniques. This is particularly relevant for understanding the observed discrepancies in health state utility gains produced by different value sets, such as for the EQ-5D-5L. We have presented a simplified 'scale adjusted unweighted' model which assigns equal weight to each dimension, as well as equal weight to each one-level change. More research is needed to develop models that account for the observed patterns of non-linearities along the steps on the level-ladder, as well as any systematic differences in the relative importance people assign to the dimensions included.

References

1. Appleby, J., Devlin, N., & Parkin, D. Using Patient Reported Outcomes to Improve Health Care. Wiley, 2016.
2. Brazier, J., Ratcliffe, J., Salomon, J. A., et al. Measuring and Valuing Health Benefits for Economic Evaluation. OUP Oxford, 2007.
3. Drummond, M. F., Sculpher, M. J., Torrance, G. W., et al. Methods for the economic evaluation of health care programme. 3rd ed.: Oxford: Oxford University Press, 2005.
4. Trauer, T., & Mackinnon, A. Why are we weighting? The role of importance ratings in quality of life measurement. Qual Life Res 2001; 10: 579-585.
5. Wisløff, T., Hagen, G., Hamidi, V., et al. Estimating QALY Gains in Applied Studies: A Review of Cost-Utility Analyses Published in 2010. Pharmacoeconomics 2014; 32: 367-375.
6. Richardson, J., McKie, J., & Bariola, E. Multi attribute utility instruments and their use. In: A. J. Culyer, ed., Encyclopedia of health economics. San Diego: Elsevier Science, 2014.
7. Torrance, G. W. Social preferences for health states: An empirical evaluation of three measurement techniques. Socio-Economic Planning Sciences 1976; 10: 129-136.
8. Torrance, G. W., & Feeny, D. Utilities and Quality-Adjusted Life Years. International Journal of Technology Assessment in Health Care 1989; 5: 559-575.
9. Parkin, D., Rice, N., & Devlin, N. Statistical analysis of EQ-5D profiles: does the use of value sets bias inference? Med Decis Making 2010; 30: 556-565.
10. Prieto, L., & Sacristán, J. What is the value of social values? The uselessness of assessing health-related quality of life through preference measures. BMC Me Res Methodol 2004; 4: 1-9.
11. Wilke, C. T., Pickard, A. S., Walton, S. M., et al. Statistical implications of utility weighted and equally weighted HRQL measures: an empirical study. Health Econ 2010; 19: 101-110.
12. Richardson, J., Iezzi, A., & Khan, M. A. Why do multi-attribute utility instruments produce different utilities: the relative importance of the descriptive systems, scale and 'micro-utility' effects. Qual Life Res 2015.
13. Brooks, R. EuroQol: the current state of play. Health Policy 1996; 37: 53-72.
14. Herdman, M., Gudex, C., Lloyd, A., et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011; 20: 1727-1736.
15. Sintonen, H. The 15D instrument of health-related quality of life: properties and applications. Annals of Medicine 2001; 33: 328-336.

16. Devlin, N., Shah, K., Feng, Y., et al. Valuing Health-Related Quality of Life: An EQ-5D-5L Value Set for England. 2016. Available from: https://www.ohe.org/publications/valuing-health-related-quality-life-eq-5d-5l-value-set-england [Accessed March 15, 2016].
17. Richardson, J., Iezzi, A., & Maxwell, A. Cross-national comparison of twelve quality of life instruments: MIC Paper 1 Background, questions, instruments. Research Paper 76. 2012. Available from: http://www.buseco.monash.edu.au/centres/che/pubs/researchpaper76.pdf [Accessed April 10, 2014].
18. Dolan, P. Modeling valuations for EuroQol health states. Medical care 1997; 35: 1095-1108.
19. Han, J., Kamber, M., & Pei, J. Data Preprocessing. In: J. H. Kamber & J. Pei, eds., Data Mining 3rd ed. Boston: Morgan Kaufmann, 2012.
20. Boyer, J. G., & Earp, J. A. The development of an instrument for assessing the quality of life of people with diabetes. Diabetes-39. Medical care 1997; 35: 440-453.
21. Jorm, A. F., Griffiths, K. M., Christensen, H., et al. Actions taken to cope with depression at different levels of severity: a community survey. Psychological medicine 2004; 34: 293-299.
22. Barnhart, H. X., Haber, M. J., & Lin, L. I. An overview on assessing agreement with continuous measurements. Journal of biopharmaceutical statistics 2007; 17: 529-569.
23. Bland, J. M., & Altman, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1: 307-310.
24. Ludbrook, J. Linear regression analysis for comparing two measurers or methods of measurement: but which regression? Clinical and experimental pharmacology & physiology 2010; 37: 692-699.
25. McGrath, C., & Bedi, R. Why are we 'weighting'? An assessment of a self-weighting approach to measuring oral health-related quality of life. Community Dentistry and Oral Epidemiology 2004; 32: 19-24.
26. Wu, C. H., Chen, L., & Tsai, Y. M. Investigating Importance Weighting of Satisfaction Scores from a Formative Model with Partial Least Squares Analysis. Soc Indic Res 2009; 90: 351-363.
27. Walters, S. J., & Brazier, J. E. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res 2005; 14: 1523-1532.
28. Alanne, S., Roine, R. P., Rasanen, P., et al. Estimating the minimum important change in the 15D scores. Qual Life Res 2015; 24: 599-606.

Table 1 – Tests for convergent validity of preference-weighted and unweighted values using Spearman rank order correlations (ρ*) for EQ-5D-5L and for 15D.

	VAS	D-39 dimensions (N=924)					D-39
	(N=7759)	EM	DC	AW	SB	SF	Average
EQ-5D-5L	0.611	0.710	0.364	0.498	0.366	0.320	0.595
Unweighted EQ-5D-5L	0.615	0.715	0.351	0.467	0.350	0.313	0.584
15D	0.665	0.766	0.433	0.524	0.400	0.397	0.660
Unweighted 15D	0.670	0.771	0.457	0.567	0.426	0.447	0.685

NB: VAS, visual analogue scale (174 missing); *EM,* energy and mobility; *DC,* diabetes control; *AW,* anxiety and worry; *SB,* social burden; *SF,* sexual functioning.
* All $p < 0.001$.

Table 2 – Tests for known group validity of preference-weighted and unweighted values for EQ-5D-5L and for 15D.

	Kruskal-Wallis H test statistics*		RE (95% CI)
	Weighted $\chi^2_{(3)}$	Unweighted $\chi^2_{(3)}$	
EQ-5D-5L			
SOL	1367.80	1301.71	1.05 (1.030, 1.071)
K10	275.07	241.07	1.14 (1.055, 1.227)
15D			
SOL	1495.03	1614.4	0.93 (0.912, 0.940)
K10	313.49	347.52	0.90 (0.869, 0.935)

NB: SOL, standard of living; *K10,* Kessler Psychological Distress Scale; *RE,* relative efficiency; *CI,* bootstrapped 95% confidence interval (with 1000 iterations); $\chi^2_{(3)}$, chi-squared statistic with 3 degrees of freedom.

* $p < 0.001$.

Table 3 – Agreement between preference-weighted *vs.* unweighted values for EQ-5D-5L and for 15D.

Measures of agreement	EQ-5D-5L		15D[a]
	Unadjusted scale [0 – 1]	Adjusted scale [-0.281 – 1]	
ICC*	0.939	0.956	0.988
[95% CI of ICC]	[0.916, 0.954]	[0.931, 0.969]	[0.988, 0.989]
Spearman's rank correlation, ρ*	0.982	0.982	0.986
[95% CI for ρ]	[0.981, 0.983]	[0.981, 0.983]	[0.985, 0.987]
Mean difference (SE)	-0.021 (0.001)	0.023 (0.001)	-0.001 (0.000)
[95% CI for mean difference]	[-0.022, -0.020]	[0.022, 0.024]	[-0.002, -0.001]
Lower limits of agreement	-0.136	-0.085	-0.038
[95% CI]	[-0.138, -0.134]	[-0.087, -0.083]	[-0.039, -0.037]
Upper limits of agreement	0.094	0.131	0.036
[95% CI]	[0.092, 0.096]	[0.129, 0.133]	[0.035, 0.037]

NB: ICC, intra-class correlation coefficient; *CI*, confidence interval; ρ (rho), Spearman's rank correlation (which is not affected by linear transformation of unweighted values); SE, standard error.

[a] No scale difference between preference-weighted and unweighted 15D.

* $p < 0.001$.

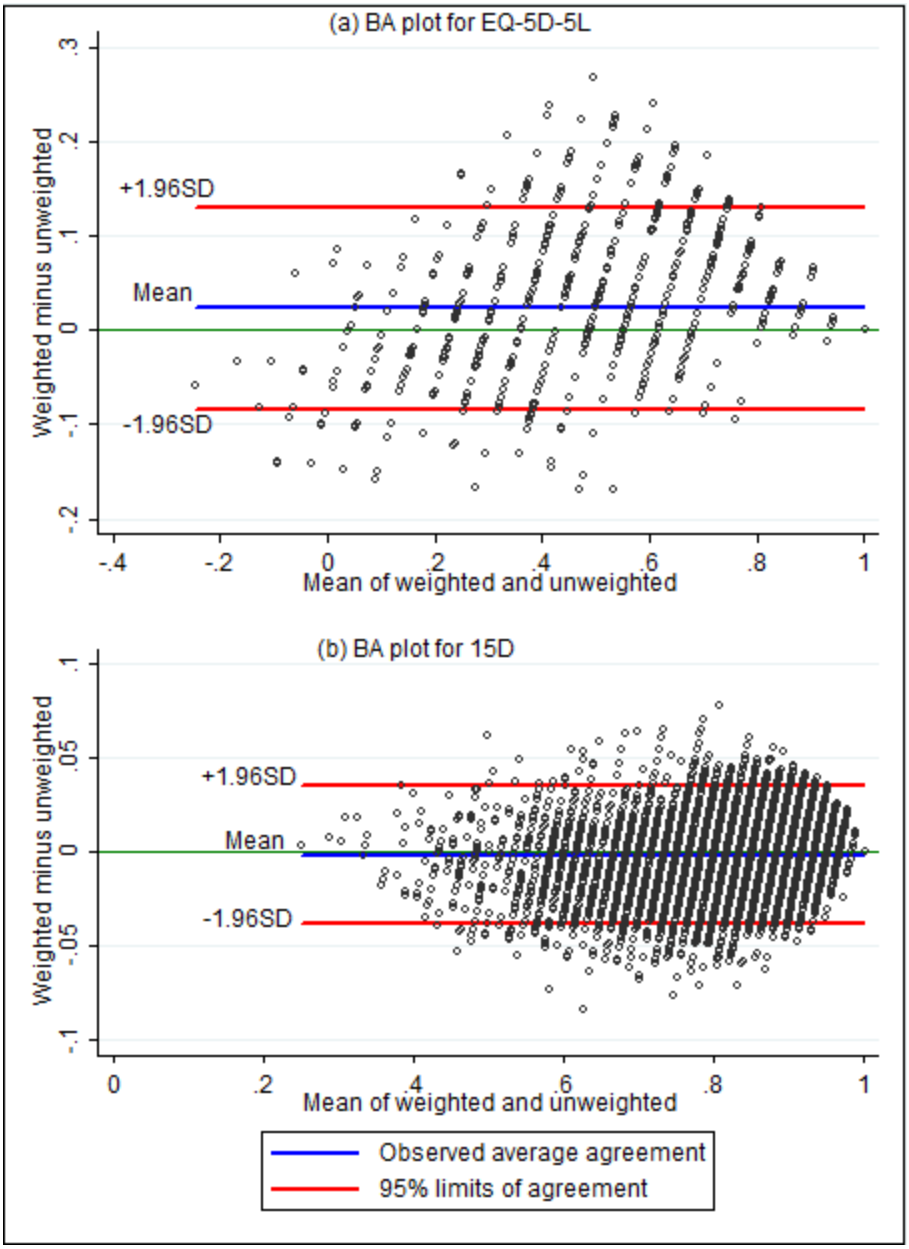

Fig. 1 – Bland-Altman plots of agreement between preference-weighted vs. unweighted values for EQ-5D-5L and for 15D.

NB: line of perfect average agreement (*green*), observed average agreement (*blue*), and the upper and lower 95% limits of agreement (*red*). Note that mean difference and the upper and lower 95% limits of agreement with 95% confidence intervals are summarized in Table 3. *BA,* Bland-Altman, and *SD,* standard deviation.

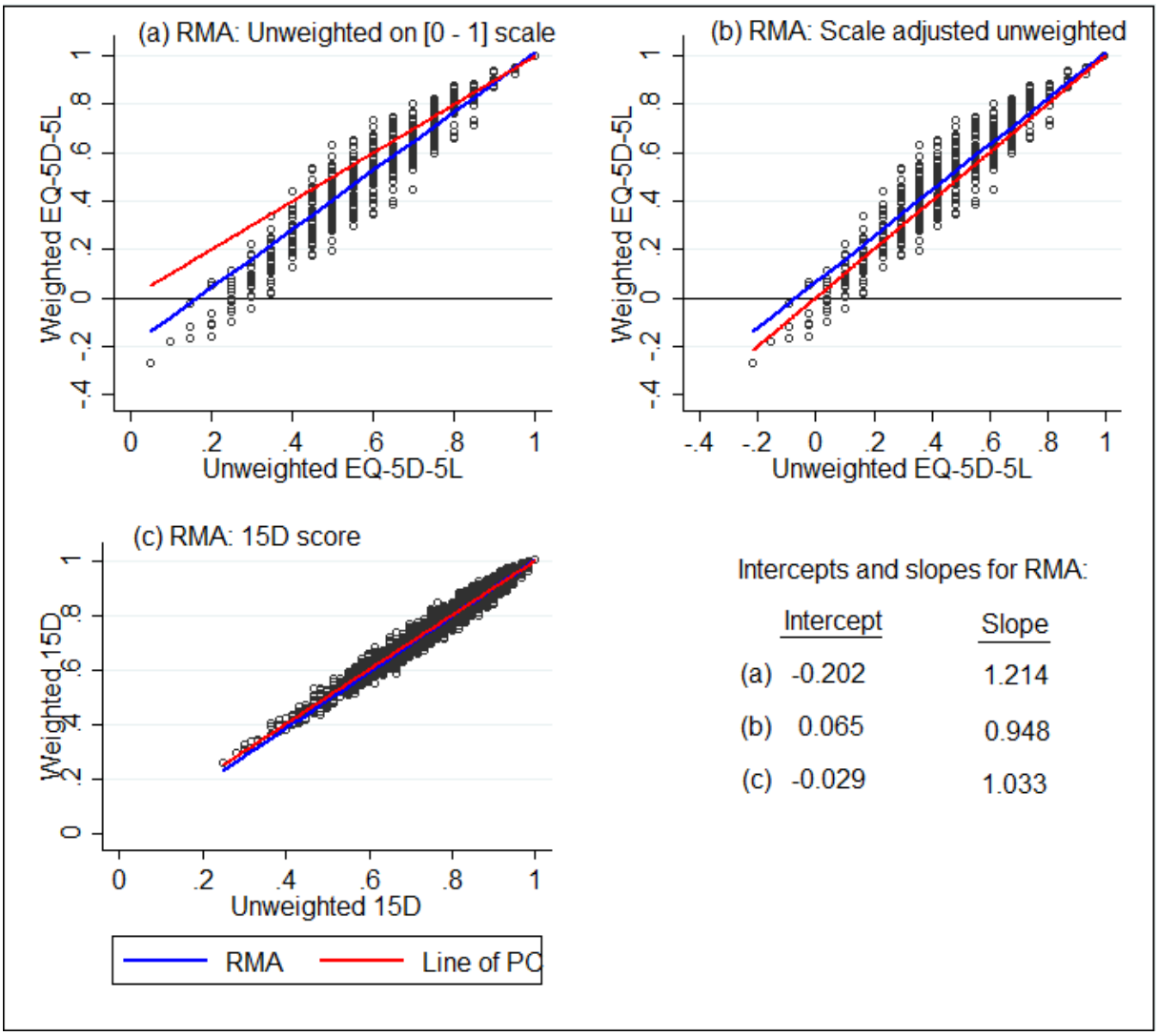

Fig. 2 – Reduced major axis (*RMA)* as a measure of agreement between preference-weighted and unweighted measures for EQ-5D-5L and for 15D.

NB: RMA (*blue*) line serves as a summary of the center of the data; *PC,* line of perfect concordance (*red*) along which preference-weighted equals unweighted values.